INFORMATION, TRADING, AND VOLATILITY: EVIDENCE FROM FIRM-SPECIFIC NEWS*

Jacob Boudoukh[a], Ronen Feldman[b], Shimon Kogan[c] and Matthew Richardson[d]

Abstract:

What moves stock prices? Prior literature concludes that the revelation of private information through trading, and not public news, is the primary driver. We revisit the question by using textual analysis to identify fundamental information in news. We find that this information accounts for 49.6% of overnight idiosyncratic volatility (vs. 12.4% during trading hours), with a considerable fraction due to days with multiple news types. As applications, we use our measure of public information arrival to reinvestigate two important contributions in the literature related to individual R^2s of stock returns on aggregate factors, namely Roll (1988) and Morck, Yeung and Yu (2000).

*Corresponding author: Shimon Kogan, Arison Business School, IDC Herzliya, Tel: +972 (9) 966-2741, email: *skogan@idc.ac.il*. We would like to thank Andrew Karolyi and two anonymous referees, John Griffin and seminar participants at the University of Texas, Austin, NYU Stern School of Business, Wharton School of Business, University of Zurich, University of Lausanne, ESMT Berlin, Case Western University, HEC Paris, York University, the 2nd Luxembourg Asset Management Summit and the discussant Guido Baltussen, participants of the 2013 WFA meetings and the discussant Paul Tetlock, as well as participants of the 2013 *AQR Insight Award* for their comments and suggestions. An earlier version of the paper was circulated under "Which News Moves Stock Prices? A Textual Analysis." We thank the Israel Science Foundation (ISF) for their financial support.

[a] Arison School of Business, IDC Herzliya.
[b] School of Business Administration, The Hebrew University.
[c] Arison School of Business, IDC Herzliya and MIT Sloan School of Management.
[d] Stern School of Business, New York University and NBER.

I. Introduction

There is a considerable literature in finance that looks at the relation between asset prices and information. Standard models in finance suggest that prices should reflect such information, whether public or private, as well as shocks to investor demand, either through liquidity shocks or irrational trading.[1] A useful tool for empirically investigating these models is the study of relative asset return variances during periods with differential information, as a way of isolating the effect of private versus public information, as well as that of noise trading.[2]

The study of French and Roll (1986) compares variance ratios of stock returns during periods of trading and overnight (i.e., non-trading hours) to better understand whether volatility is caused by public information, private information (revealed through trading), or pricing errors by investors. Their rationale is that private information can affect volatility only during trading hours as information is gradually revealed through trading. French and Roll (1986) conclude that private-information driving rational trading is the main driver of return volatility. Using more complete data and additional real-world experiments, this conclusion has generally been confirmed by later studies.[3]

An alternative view of the evidence has been expressed by the burgeoning literature in behavioral finance. For example, Hirshleifer (2001) writes "little of stock price variability has been explained empirically by relevant public news," (p.1560); Shleifer (2000) writes "movements in prices of individual stocks are largely unaccounted for by public news..."; and Hong and Stein (2003) write "Roll (1984, 1988) and French and Roll (1986)

[1] See, for example, Grossman and Stiglitz (1980), Milgrom and Stokey (1982), Kyle (1985), Tauchen and Pitts (1983), and Glosten and Milgrom (1985), among early papers in the field. For the liquidity shock channel, see Admati and Pfleiderer (1988) and Foster and Viswanathan (1990), and, for irrational trading, see Black (1986), De Long, Shleifer, Summers and Waldmann (1990), and Daniel, Hirshleifer and Subrahmanyam (1998).

[2] See, for example, French and Roll (1986), Roll (1988), Barclay, Litzenberger and Wharton (1990), Francis, Pagach and Stephan (1992), Green and Watts (1996), Jones, Kaul and Lipson (1994), Jiang, Likitapiwat, and McInish (2000), Fleming, Kirby and Ostidek (2006), Cliff, Cooper and Gulen (2008), Kelly and Clark (2011), and Lou, Polk and Skouras (2018), among others.

[3] See, for example, Barclay, Litzenberger and Wharton (1990), Madhavan, Richardson and Roomans (1997), Ito, Lyons and Melvin (1998), Barclay and Hendershott (2003), and Chordia, Roll and Subrahmanyam (2011).

demonstrate in various ways that it is hard to explain asset price movements with tangible public information" (p.487).

Our contribution is to show that firm-level public news, which we refer to as "news" henceforth, is a meaningful component of stock return variance. Using textual analysis, we identify from news stories *relevant* public information tied to specific firm events. We then re-evaluate existing findings by identifying informationally relevant non-trading and trading periods, thus controlling for private information induced volatility.[4]

Common to much of this literature, the proxy for public information has been news articles.[5] A problem with this proxy is its potentially low power. Common news sources for companies, such as those in websites, the Wall Street Journal stories, and Dow Jones News Service, release many articles that may contain very little relevant information about company fundamentals. The goal for the researcher is to be able to parse through news stories and determine which are relevant and which are not. However, combing through hundreds of thousands, possibly millions, of news stories is a massive task. Fortunately, advances in textual analysis allow for better identification of relevant news. This paper employs two independent approaches that systematically and objectively identify events within news articles: (i) a commercially-available information extraction platform, called *Visual Information Extraction Platform* (*VIP*), and (ii) a machine-learning method that is an industry standard, *Ravenpack*. The paper focuses on the first approach, as we are able to make the data for it publicly available, while replicating the main findings with the second approach.[6]

Using these two approaches, we match each news article, which itself is time-stamped and linked with stock ticker(s), to a series of either "identified" value-relevant events or

[4]Other papers focused on informationally relevant periods include Jones, Kaul, and Lipson (1994), Fleming, Kirby and Ostdiek (2006) and Jiang, Likitapiwat, and McInish (2012). For example, Jones, Kaul and Lipson (1994) investigate individual stock return volatility on trading days with no volume; Jiang, Likitapiwat and McInish (2012) study after-hours trading when earnings are released overnight; and Fleming, Kirby and Ostdiek (2006) analyze commodity return volatility between trading and overnight hours during periods when prices are theoretically more sensitive to weather.

[5] See, for example, Roll (1988), Chan (2003), and Tetlock (2007).

[6] For more details on both commercial platforms, see http://www.amenityanalytics.com/ and https://www.ravenpack.com/. The *VIP* dataset employed in this study is available here https://www.dropbox.com/s/ilw2l2pb9nh6kpm/InformationTradingVolatility_DataFile.csv?dl=0.

"unidentified" news. In particular, we examine stock return variation during trading hours and overnight around specific types of news such as unidentified news (news with no identified, value-relevant, topic), identified news (news with an identified, value-relevant topic), and identified news with different levels of complexity (to be defined precisely later on). As a proof of concept, we document that stock-level volatility varies greatly with the type of news – identified or unidentified – but not so much with the presence of news. On identified (complex) news days, the variance of stock prices is more than twice (four times) that of other days, consistent with the idea that the intensity and importance of information arrival is not the same across these days.

Using our identification of relevant news, we revisit some of the key analysis of French and Roll (1986). Notably, we find a large difference in the change in volatility on news days when comparing trading hours versus overnight. Because overnight returns are largely unaffected by private information-driven trading, we are better able to identify volatility arising from public information. Doing so, we find that the overnight variance ratio of identified (complex) news to no news is 4.76 (10.11), a magnitude higher than the same ratio during trading hours, namely 1.78 (2.42).

Consistent with the findings of French and Roll (1986) we find that unconditional median return volatility during trading hours is 2.02%, 68% higher than overnight volatility, namely, 1.20%. In contrast, the median volatility conditional on identified news is only 20% higher during the trading day relative to overnight (2.51% versus 2.09%), and actually 3% lower during complex news days (3.01% versus 3.10%). When conditioning on identified news, these findings provide a contrast to conclusions reached by French and Roll (1986) and others who document considerable more volatility during trading hours unconditionally.

A key contribution of this paper is to provide a methodology that allows us to isolate the portion of return variance due solely to relevant news. The underlying assumption is that there is a continuous stream of unexplained return variability, so that even on days with important news, part of the return variability is untied to that news, thus providing a lower bound on the variance contribution of public information. We apply the methodology to overnight and trading hours, conditional on different types of information (such as unidentified or no news, identified news, complex news, and specific events) and compare

the results from this decomposition. Intuitively, we decompose the contribution of news to overall return variance into the intensity of news arrival and the impact of news conditional on news arrival. We also differentiate between market-wide and idiosyncratic volatility. We show that, for S&P500 firms, idiosyncratic variance explained by public information is around 49.6% of total overnight variance and around 12.4% during trading hours. Even though only 27% (22%) of news days are complex news days during overnight (trading) hours, 32.4% (6.1%) of all the idiosyncratic variance is explained. The variance contributions from all news days are higher for firms that have higher trading volume, i.e., 57.0% overnight and 14.15% during trading hours. Moreover, particular types of events have on average differential impacts on volatility, with, not surprisingly, financial-related events being the most important. That said, while these events, along with mergers, tend to be more likely, less common information such as news about the ratings of the company or financing have a large effect.

The bottom line of the paper is that, when relevant news can be identified, stock price movements are closely linked to the arrival of this information. We leverage off this finding and explore two studies related to individual R^2s of stock returns on aggregate factors, namely Roll (1988) and Morck, Yeung and Yu (2000). We ask whether better identification of firm-specific information provides insights to key findings in these papers.

First, Roll (1988) documents similar market model R^2s on news and no-news days, which places a challenge to researchers trying to explain stock price movements, e.g., in our sample, R^2s are 34.5% and 33.9% respectively on no news and unidentified news days. In contrast to Roll (1988) but consistent with theory, the R^2s are much lower on identified news days, i.e., 17.7%, and especially so on complex news days, i.e., 10.0%. While these results partially resolve Roll's R^2 puzzle, there are not enough identified news events (and aggregate movements) to explain stock return volatility. Using a measure of unexplained volatility, we cross-sectionally confirm an implication from the noise trading literature, namely that higher noise trading (as proxied by a lower fraction of news variance contribution) is associated with higher expected returns.

Second, Morck, Yeung and Yu (2000) document declining market model R^2s for individual firm stock returns from 1926 to 2000, yet in a recent paper Morck, Yeung and Wu (2013)

report the downward trend of R^2s has reversed. Appealing to work by Veldkamp (2006a, 2006b) on endogenous information production, we argue that the benefit to private information production has decreased due to the increase in publicly available information. To this point, we document a large negative correlation over time (i.e., -0.50) between average idiosyncratic volatility during overnight hours and the contribution of identified news to overnight return volatility. Additional empirical evidence, such as the properties of volatility conditional on identified news versus unidentified news, support the hypothesis that private information's value may have decreased.

II. Data Description

A. Textual Analysis

Financial economists use data to help evaluate theories of financial behavior. The vast majority of data is inherently unstructured – corporate filings and announcements, financial news, professional financial analysts reports, and so forth – and therefore difficult to use in practice. Over the last decade, with the increase in computing power and collection of massive amounts of data, the field of finance has made large inroads into text mining. See, for example, Gentzkow, Kelly and Taddy (2017) for a recent survey of historical advances and recent innovations in textual analysis in the social sciences with an emphasis on finance.[7]

A particular application in finance has been to use textual analysis to help forecast short-term stock price movements. The preferred methodology has focused on word counts based on dictionary-defined positive versus negative words. For example, one of the prominent papers is Tetlock (2007). Tetlock (2007) employs the *General Inquirer*, a well-known textual analysis program, alongside the *Harvard IV-4* dictionary, to calculate the fraction of negative words in the *Abreast of the Market* Wall Street Journal column.[8] Statistical-based methods for interpreting text, including supervised machine-learning techniques, have also

[7] See also Feldman and Sanger (2006) for a general analysis of text mining.

[8] Tetlock (2007) gave rise to a large number of papers that apply a similar methodology to measure the positive versus negative tone of news for forecasting stock returns (e.g., Tetlock, Saar-Tsechansky and Macskassy (2008), Bollen, Mao and Zeng (2011), Loughran and McDonald (2011), Garcia (2013), Wisniewski and Lambe (2013) and Chen, De, Hu and Hwang (2014), among others).

been employed to forecast stock returns. Supervised learning involves the researcher fitting a model to a training data set in which the outcome (e.g., the news sentiment) has been pre-determined, and then applying the results to a much broader data set for testing.[9] The earliest example in finance is Antweiler and Frank (2005) who apply language algorithms to analyze internet stock message boards posted on "Yahoo Finance" to help predict stock return movements.[10]

A growing literature in finance and accounting now uses textual analysis to measure the content and tone of documents for applications beyond stock market prediction.[11] While this paper also uses textual analysis, the focus of this paper is quite different. We are not interested in identifying the sentiment of the news *per se*, but rather its saliency. Specifically, we use textual analysis to identify events relevant to companies, such as new product launches, lawsuits, analyst coverage, news on financial results, mergers, *et cetera*. While stock tickers and words associated with specific events in a given article are fairly straightforward to identify, linking a particular company to a specific event in that same article is not trivial. We use two independent methodologies from separate commercial platforms, both based on supervised learning, to identify these events. Supervised learning is particularly appropriate for this application because an extensive, highly accurate training data set of company events in news articles can be created.

The first approach, *Visual Information Extraction Platform (VIP)*, uses a mixture of a rule-based information extraction platform and a trained support vector machine classifier in order to identify event instances for companies and measure sentiment from text contained

[9] Unsupervised learning methods contrast with supervised learning ones to the extent the outcome (e.g., the news sentiment) is not observed and instead must be inferred from some assumed structure. For a recent example of this approach to economics, see Hansen, McMahon, and Prat (2014) who study the transcripts of FOMC meetings.

[10] See also Das and Chen (2007), Jegadeesh and Wu (2013) and Heston and Sinha (2014).

[11] See, for example, Li (2008), Feldman, Govindaraj, Livnat, and Segal (2010), Davis, Piger and Sedor (2011), Tetlock (2011) and Loughran and McDonald (2014) who use a dictionary-based approach and Li (2010), Hanley and Hoberg (2011), Hoberg and Phiips (2010, 2016), Grob-Klubmann and Hautsch (2011), Kogan, Routledge, Sagi and Smith (2011), and Manela and Moreira (2017) who employ machine learning-based applications to decipher either the content or sentiment of the text.

in financial news.[12] We apply *VIP* to Dow Jones Newswire because we want to focus on breaking news about the firm that matches a specific date and time. (Of course, some of the news might be analysis from the Dow Jones newsroom and may not be a breaking story.) For this reason, we purposefully do not expand to other media outlets and websites. *VIP* conditions on firm-specific events and thus parses out which news items are more likely to be relevant for firm valuation. These events include those covered by Capital-IQ and reported in a cross-section of academic event studies. There are ninety event subcategories which are further categorized into one of 18 categories: *Business Trends, CSR Brand, Capital Returns, Deals, Earnings Factors, Employment, Facility, Financial, Financing, Forecast, General, Investment, Legal, Mergers & Acquisitions, Product, Ratings, Stock and Stock Holdings*. (See Appendix A for a comprehensive list.)

To better understand how the methodology works, a random training sample of 2,000 articles on S&P500 firms were chosen and then read by finance-knowledgeable readers. Approximately half the articles could be matched to the set of event types. The sample of these news days were then used as training data by the *VIP* textual analysis classifier. After training on the smaller sample, the *VIP* classifier was then applied to the approximate 840,000 stock/day observations with news (many of which contain multiple articles) on S&P500 firms from 2000-2015.

Specifically, for each of the articles from the training sample, *VIP* looks for words and phrases connected to the event subcategories. For example, consider the *Employment* category. Changes in the CEO, an executive of the firm or a board member; executive compensation; and employment issues including, among other items, strikes and changes in the workforce are all flagged. Sentences involving these words or phrases are trained on a support vector machine classifier given *VIP* knows which article is truly about the event. Across all the articles, VIP then uses the support vector machine model to classify each sentence involving the relevant words and phrases into one of the events or possibly no

[12] As is the case with other commercial platforms (such as *RavenPack* also used in this study), much of the interest is related to measuring sentiment. This is not a focus of this paper as our use of these platforms is in identifying whether a piece of news is potentially relevant for the firm's valuation.

event based on its cosine similarity with the "identifying" versus "non-identifying" sentences.[13] The machine learning procedure therefore cuts down on false positives.

A significant difficulty arises because candidate sentences that may contain events often do not mention the specific name or ticker of the company which is the subject of the sentence. The methodology underlying *VIP* helps resolve these indirect references by analyzing the flow of the article and by utilizing anaphora resolution techniques[14]. In particular, using co-reference resolution techniques (Lee, Chang, Peirsman, Chambers, Surdeanu and Jurafsky (2013)), *VIP* applies rules to relate topic sentences to the last active company name or to its indirect reference such as "the company", "the firm", "it" etc. The benefit of the classifier therefore is that it allows the researcher to determine whether a significant event occurred for a particular firm at a specific time of day across *all* Dow Jones Newswire stories. Of potential interest to researchers, the online dataset includes the ticker-event-date dataset used in this paper (full day, open to close, and close to open).[15]

The second methodology comes from *RavenPack, Inc*., which represents an industry standard for asset management firms in terms of news analytics. *Ravenpack* uses machine learning algorithms to process text from not only the Dow Jones newswire but also the *Wall Street Journal*, direct regulatory feeds, and thousands of social media websites, into a machine-readable content to identify a company's news in terms of "relevance" and a "sentiment". Specifically, every time a company is reported in the news, *RavenPack* produces 16 fields such as a time stamp, company identifiers, scores for relevance, novelty and sentiment, and unique identifiers for each news story. Because *RavenPack* casts a wide net, many of the news stories that link to a stock ticker are not relevant for the firm's valuation. As a result, *RavenPack* provides a relevance score from 0 to 100. While this score is a black box, a relevance score of 100 generally coincides with the company playing

[13] For a discussion of cosine similarity and its use in supervised learning in text processing, see, for example, Manning, Raghavan and Schütze (2008).
[14] Anaphora resolution, a key issue with natural language processing, is the problem of resolving references to earlier or later items in the discourse (see Mitkov (1999) for a survey).
[15] The *VIP* dataset can be downloaded here: https://www.dropbox.com/s/ilw2l2pb9nh6kpm/InformationTradingVolatility_DataFile.csv?dl=0.

a main role in the story and the article type being identified. *RavenPack* itself recommends a score of at least 90 (and possibly 100) to identify relevance.

While both methodologies apply different approaches to different data - *VIP* uses the Dow Jones newswire while *RavenPack* uses more sources albeit with a relevance score - each database contains a unique observation for every article and includes a time stamp plus a number of variables that identify the content and form of the article. Perhaps not surprisingly, the results using both methods are qualitatively similar. Because the specific event is identified under *VIP* and made available for researchers, for the remainder of the paper, we document results using the *VIP* identification of specific events. That said, for two of the key tables, Appendix B duplicate the findings using *RavenPack*.

B. Data Set Construction

As described above, not all days a firm is mentioned in a news story are relevant in terms of breaking news. We want to condition on firm-specific events which are more likely to be relevant for firm valuation. The primary dataset used in this paper consists of all documents that pass through the Dow Jones Newswire from January 1, 2000 to December 31, 2015. For computational reasons, and to minimize issues related to poor tradability, we limit ourselves to S&P500 companies with at least 20 trading days during the period. Over the sample period, the dataset therefore includes a total of 896 companies. To avoid survivorship bias, we include in the analysis all stocks in the index as of the first trading day of each year. We obtain price and return data from CRSP.

To ensure that the analysis does not suffer from a look-ahead bias, we use the article timestamp and line it up with the trading day. Specifically, we consider as date *t* articles those that were released between 16:00 on date *t-1* and 16:00 on date *t*. Date *t* returns are computed using closing prices on dates *t-1* and *t*. We also perform an analysis using trading hours (open-to-close) and overnight (close-to-open) returns. For these returns, open-close news is defined as news arriving during trading hours and close-open news is defined as

news arriving after trading hours.[16] Articles released overnight (weekends and holidays) are matched with the next available trading day. *VIP* methodology processes each article separately and generates an output file in which each article/stock/time stamp is represented as an observation.

For each of the aforementioned observations, *VIP* reports the total number of words in the article, the number of relevant words in the article, and any possible identified events (and sub-events) as described in II.A above. A key feature of the methodology is its ability to differentiate between relevant news for companies (defined in our context as those related to specific firm events) as opposed to unidentified firm events. For each news story, therefore, our application of *VIP* produces a list of relevant events connected to this company and to this particular piece of news. Note that multiple events may be connected to a given story.

As shown below, *VIP* and *Ravenpack*'s event identification are successful to the extent stock prices move much more on event days than unidentified news days. One possible explanation is reverse causality, that is, once stock prices move, the media or analysts make up the news. While clearly this can take place, note that the identified news are specific events related to the firm as described in Appendix A. The specificity of the event creates a high hurdle for manufacturing news. Further, a number of days with identified news have low volatility, while some days with no or unidentified news have high volatility. As a final comment, for some of the applications to follow, different return behavior is produced, conditional on whether the news is identified compared to the size of the stock price move. That said, the issue of reverse causality in text processing in finance deserves greater scrutiny.

This issue aside, our goal is to analyze the difference in return patterns based on the type of information arrival. We therefore classify each stock/period into one of three types:

[16] In this paper, we define as trading hours by the hours during which the stock exchange is open. Trading also takes place during overnight hours, rising from approximately 5% trading volume in the early part of the sample to as much as 15% in the latter part. Whether this level of volume is sufficient to induce private-information based trading is an open question though Jiang, Likitapiwat and McInish (2012) analyze after-hours trading around earning announcements and document informational efficiency during these hours.

1. *No news* – observations without news coverage.
2. *Unidentified news* – observations for which none of the news coverage is identified by one of the eighteen categories.[17]
3. *Identified news* – observations for which at least some of the news coverage is identified as being related to one of the above categories.

As mentioned in Section II.A above, we limit our analysis to the Dow Jones Newswire in an attempt to capture breaking news that is more likely to contain new and relevant information. News on other media sites may also contain unique information which our methodology will miss if this news does not appear on the Dow Jones Newswire (or appears much later). The impact of this assumption will be to reduce the relative power of our identified news measure. While arguably less an issue for the large firms (S&P500 constituents) studied here, extensions of our methodology to small firms, with a more local following, deserves greater examination.

Nevertheless, conditional on being classified as *identified news*, we provide a further breakdown of identified news, with a subset of these days being denoted as *complex news* days, defined as *identified news* days with more than two events (either categories or subcategories). The motivation for separating out multiple event days is twofold. On the one hand, there is a body of the behavioral finance literature that argues and documents investors have difficulty interpreting complex news and firms (e.g., Barber and Odean (2008), DellaVigna and Pollet (2006) and Cohen and Lou (2012)). Our multiple event days can be treated as such in that context. On the other hand, these complex days may simply be associated with more salient news, so these multiple event days represent more intense and relevant news days than single events days. Interestingly, as a preview, the latter interpretation seems to hold as complex news days tend to be more important in terms of variance contribution.

[17] The *RavenPack* database provides a "relevance" score ranging from 0 to 100 for each news story. This score is a measure of how closely a company is related to the specific event underlying the story. In the analysis to follow, we denote scores of "100" as identified news, and all other scores associated with news as unidentified. Under this classification, *RavenPack* identifies "relevant" news on par with *VIP* (e.g., 17.2% versus 21.6% of all days).

In addition, we consider three periods covering news and returns:

1. *DAY* – the full trading day, including trading hours and the night hours during which the market is closed, until the start of trading the next day.
2. *TRDNG* – trading hours (from open to close).
3. *OVRNT* – overnight (from close to open).

The news story may be stale in terms of the event's timing, so that the event occurred sometime earlier. In that case, using the time-stamp on the news article as a proxy for when the event occurred would reduce our power to link information with price movements.

Figure 1 and Table 1 provide an overview of the data. For each hour over the 24-hour period, Figure 1 provides the average number of S&P500 firms that have news stories, either identified by a specific event or unidentified. While much of the news (both identified and unidentified) hovers around trading hours, the peak of the news events occurs in the hour or two just before and after trading hours. This finding is consistent with Bradley, Clark, Lee and Ornthanalai (2014) who document that analyst upgrades and earnings announcements primarily occur outside the trading hours.

To be more precise, the first column in panel A of Table 1 reports the number of observations under each of the day classifications documented by *VIP*. The majority of days have no news coverage, i.e., 1,112,341 of 1,952,175 (or 57.0%) stock/day observations contain no news reported on the Dow Jones Newswire[18]. There are 839,834 news days, out of which 417,889 (or 49.8%) days do not have an identified topic news event. Of the 421,945 identified news days, only 124,824 (or 29.6%) are complex news days. Table 1 also observes that identified news days contain a larger number of articles compared with unidentified news days (4.9 vs. 1.9 per stock/day). While the number of words per article does not seem to vary much by day type, the number of relevant words (as identified by *VIP*) is much larger on identified news days (119 vs. 70).

[18] As a comparison, the last column of Table A.2 in the Appendix reports the number of observations under each day classification using *Ravenpack* data. *Ravenpack* casts a much wider net in terms of news events as only 21.76% of days have no coverage. However, of some importance, the vast majority of the days with news coverage in *Ravenpack*, 78.0%, have a relevance score less than 100% and are considered unidentified.

Panels B and C of Table 1 report similar statistics but now broken down between trading hours and overnight. As mentioned above, the most interesting result is that, while the number of unidentified events is similar, more relevant news coverage occurs during periods when the market is closed versus open. For example, the ratio of news days - unidentified, identified and complex - to total number of days - is respectively 15.1%, 11.4% and 2.5% during trading hours versus 15.9%, 15.4%, and 4.2% overnight.[19] While this finding may have something to do with when news crosses the "wire", as opposed to when it takes place, it nevertheless suggests a continual volume of news throughout a day, with perhaps more value-relevant news occurring after the close of trading (see, for example, Bradley, Clark, Lee and Ornthanalai (2014)). This fact will be useful when the return distributions are compared across different types of news.

III. **Return Volatility and News**

A basic tenet of financial economics is that asset prices change in response to unexpected fundamental information. Section II.B describes a wide variety of news types from unidentified to identified to identified complex news. What differential impact does this news assortment have on the distributional properties of returns? Identifying which news is relevant is important because a number of empirical results in the literature depend on showing that the distributional properties of stock prices are similar on news versus no news periods.

Early work, primarily though event studies, seemed to confirm a strong link between prices and specific events. (See, for example, Ball and Brown (1968) on earning announcements, Fama, Fisher, Jensen and Roll (1969) on stock splits, Mandelker (1974) on mergers, Aharony and Swary (1980) on dividend changes, and Asquith and Mullins (1986) on common stock issuance, among many others.) However, since Roll's (1988) provocative presidential address showing little relation between stock prices and news (used as a proxy for information), the finance literature has provided many analyses which demonstrate little

[19] *Ravenpack* provides similar findings to the extent that identified news events are more common during the overnight hours than during the trading day, with unidentified news events being equally likely. As mentioned above, however, *Ravenpack* captures many more news days (albeit with less importance) once their relevance score is relaxed.

relationship between prices and news, e.g., see Shiller (1981), Cutler, Poterba and Summers (1989), Campbell (1991), Berry and Howe (1994), Mitchell and Mulherin (1994), and Tetlock (2007), to name a few.[20] The conclusion from this literature is that stock price movements are largely described by irrational noise trading or through the revelation of private information through trading. As pointed out in Section II, however, one of the issues with this literature may be the inability to recognize which news is relevant or not.

In this section, we document the properties of stock return variance ratios with that of identifying periods of relevant news. The first analysis we perform is a simple comparison of variance ratios of stock returns during periods with different amounts of relevant news. As a first pass at the data, Table 2 provides a breakdown of news stories by the distribution of returns. If identified news days proxy for information arrival, then we should find that news arrival would be concentrated on days with large return movements, positive or negative. To relate news arrival intensity with returns, we assign daily returns into percentiles separately for each stock and year: bottom/top 10% (i.e., the extreme 20% of returns), moderate 40% of return moves, and the smallest 40% return moves. We perform the assignment for each stock separately to control for cross-sectional variation in total return volatility, and perform the assignment for each year separately to control for systematic time-series variations in average return volatility, e.g., 2008-9. The columns in Table 2 group observations according to this split. For each of these columns, we compare the observed intensity of different day types to the intensity predicted under the null that these distributions are independent. The results in each row report the difference between the observed intensity and the null in percentage terms.

Table 2A reports the results for daily returns. First, we find that no news days are less concentrated among days with large price changes. They are 7.9% less likely to be extreme

[20] Some recent exceptions are Griffin, Hirschey and Kelly (2011), Engle, Hansen and Lunde (2011) and Neuhierl, Scherbina and Schlusche (2013). While the focus of each of these papers is different, these papers provide some evidence that better information processing allows researchers to establish a stronger relation between prices and news.

relative to the unconditional. Interestingly though, we observe very little evidence of extreme price changes on news days when we cannot identify a specific event tied to the news: only 0.5% more than the expected fraction of our defined "extreme" days. *Ex-ante*, one might have imagined that large price moves would have generated "news" stories, but this result shows no mechanical relation between news and firm volatility. Second, in sharp contrast to these results, identified news days are 24.9% more likely to coincide with the bottom 10% and top 10% of return days. That is, identified news days are much more likely to be extreme return days. Third, this pattern is much more pronounced for complex news days; these days are 63.3% more likely to coincide with extreme returns days. This finding provides some evidence that investors recognize the importance of these days.

In line with the existing literature, we study the link between news arrival and volatility by computing daily return variations on no news days, unidentified news days, identified news days and complex news days. Specifically, for each stock we compute the average of squared daily returns on these day types. We then calculate the ratio of squared returns on unidentified news days to no news days, and the ratio of squared returns on different types of identified news days to no news days.[21] For example, if both unidentified and identified news days have no additional effect on stock volatility, then these ratios should be distributed around one.

The last three columns of Table 2A report the distribution of these variance ratios. Consistent with the aforementioned results, we find that the median variance ratio of unidentified news days to no news days is close to one (i.e., 1.16) while the variance ratio of identified news days exceeds two (i.e., 2.17). The result appears quite robust, with over 90% of stocks exhibiting variance ratios exceeding one on identified news days. These results are much larger for complex news days, with 4.23 times the variance ratio. As additional evidence, Figure 2 depicts the distribution of these ratios across the 832 stocks for which these ratios are available (out of 896), winsorized at 10.[22] As evident, the ratios are not distributed around one for either unidentified or identified news days. However, the

[21] We include only stocks with at least 20 observations for all day classifications.
[22] Here again we eliminate stocks with insufficiently many observations in each day type, similarly to the footnote above.

difference in distributions between unidentified and identified news days' ratios is clear: the variance ratio is much higher on identified news days compared with unidentified news days. Note that Table A.2 in the Appendix provides a similar analysis to Table 2A using the *Ravenpack* data source, confirming the above findings.

A. Variance Ratios During Trading Hours and Overnight

The results above clearly demonstrate that the news classification procedure has power to distinguish between days on which price-relevant information arrives. This subsection uses this news classification to revisit some well-known conclusions about the predominant role of private information arrival on stock return volatility.

French and Roll (1986) compute variance ratios of stock returns during trading hours and overnight to study the role of trading on return volatility. They document considerably more variability of returns during trading hours than overnight both on an absolute and hourly basis. French and Roll (1986) explore three possible explanations. First, public information may arrive more frequently during trading hours. They provide evidence against this hypothesis by showing that volatility drops over weekday exchange holidays when presumably information is still flowing. Complementary to this finding, Table 1, Panels B and C of this paper show that identified, i.e., relevant, news seems to be generated a little less during trading hours than overnight (i.e., 11.4% and 2.5% for identified and complex news versus 15.4% and 4.2%, respectively).

Second, appealing to behavioral finance, trading itself may generate noise and thus higher volatility. Supply and demand shocks, possibly weakly related to fundamentals, affect prices through elastic supply and demand curves. Third, private information, not public information, may be the primary source for volatility. That is, private information is gradually revealed through trading, thus generating higher volatility during trading hours. French and Roll conclude that the evidence favors the latter channel and strongly supports private-information, rational, trading models. (See also Barclay, Lizenberger and Warner (1990), Madhavan, Richardson and Roomans (1997), Ito, Lyons and Melvin (1998), and Barclay and Hendershott (2003), to name a few).

As described in the introduction, a number of papers compare return variances during trading hours and overnight as a way of isolating relevant information (e.g., Jones, Kaul, and Lipson (1994), Fleming, Kirby and Ostdiek (2006), and Jiang, Likitapiwat, and McInish (2012)). These papers document that significant volatility occurs overnight, concluding that public information is an important component of price variability. Consistent with these studies, in this subsection, we reexamine the results of Table 2A, but now break the returns and news type data into trading hours and overnight. Specifically, Table 2, Panels B and C (and Table A.2, panels B and C in the Appendix) compare variance ratios of stock returns on unidentified news, identified news, and complex news days to no news days, conditional on trading versus no trading hours.

With respect to the existing literature on stock return variances during trading hours versus overnight, Table 2, panels B and C, confirms the stylized fact on variance ratios for S&P500 firms – the median trading hours daily return volatility is 2.02% versus 1.20% overnight, that is, 68% higher. On the surface, this result is consistent with the conclusions in French and Roll (1986) and others that the major source for return volatility is not public information, but instead either private information revealed by trading or noise trading. Moreover, return variances are relatively higher during trading days with no news, that is, on days without discernible public information. Specifically, median return volatility during trading hours versus overnight is 1.90% versus 0.92%, respectively; that is, 107% higher on no-news days.

Tables 2 Panels B and C, however, reveal a different story when the news can be identified, and especially so when the news is complex. Specifically, on identified news days, the median trading day volatility is 2.51% versus overnight volatility of 2.09%, in other words, only 20% higher on identified versus 68% for unidentified news days. Equally important, the identified news median volatility of 2.09% overnight is higher in magnitude than the volatility during trading hours on no-news days (i.e., 1.90%). This latter result is important for understanding the source of volatility and illustrates the importance of public versus private information in explaining return variability. These results are even stronger for complex news. In particular, overnight volatility is similar to trading hours volatility, i.e.,

3.10% versus 3.01%. These results suggest that public information, when appropriately identified, is a much more important source of volatility than previously considered.

A corollary of these findings relates to return variances conditional on various news types, both overnight and during trading hours. Specifically, overnight, the median variance ratio of returns on unidentified news, identified news and complex news days to no news days is 1.26, 4.76, and 10.11, respectively. This contrasts with significantly lower variance ratios during trading hours, i.e., 1.25, 1.78, and 2.42, respectively, for the various news types.

On the one hand, this result supports the idea that private information (or noise trading) is an important determinant of stock return volatility. That is, variance ratios are lower during trading hours when private information can be revealed through trading in contrast to overnight. Of course, as described in footnote 17, trading also occurs during overnight hours. This trend towards overnight trading has increased over our sample period albeit still at low volume levels (e.g., Jiang, Likitapiwat and McInish (2012)). Nevertheless, as a first approximation, the bifurcation between trading and overnight hours seems to capture important differences. On the other hand, on identified and complex news days, the variances are 78% to 142% higher than no news days, even during trading hours. That is, when one can identify relevant information, this information clearly plays an important role in explaining stock return volatility. This finding is amplified overnight, when much less trading takes place during overnight hours. Overnight, the stock return variances are 376% to 900% higher on identified news and complex news days relative to no news days. We confirm these results qualitatively using *Ravenpack* data, reported in the Appendix, Table A.2, panels B and C.

IV. **Return Variance Decomposition**

Section 3 presents overwhelming evidence that (i) there is greater return variation on days with specific news events, and (ii) this greater return variation diverges depending on whether the news is released during trading or overnight. The evidence, however, does not quantify how important news are for *overall* return variability. Our goal is to quantify the contribution of public information to stock return volatility. We cannot simply estimate the return volatility on days with public information because return volatility exists on days

without any relevant information. In this section, we suggest a simple model that lets us decompose total return variance into return variances that is due to private information, public information, and noise. In addition, we separate out stock return variance into its systematic and idiosyncratic component to further isolate the impact of news.

Assuming no serial correlation, daily return volatility can be broken up into two components: trading hour return volatility and overnight return volatility. (We address this assumption below.) These returns can further be separated into components conditional on identified news versus no news/unidentified news days, that is,

$$\begin{aligned}\sigma^2_{DAY,jt} \approx\ & p_{OVRNT:News,jt}\sigma^2_{OVRNT:News,jt} + p_{OVRNT:NoNews,jt}\sigma^2_{OVRNT:NoNews,jt} \\ & + p_{TRDNG:News,jt}\sigma^2_{TRDNG:News,jt} + p_{TRDNG:NoNews,jt}\sigma^2_{TRDNG:NoNews,jt}\end{aligned} \quad (1)$$

where $\sigma^2_{DAY,jt}$ is the daily return variance of firm j at time t; $\sigma^2_{OVRNT:News,jt}$ is the overnight return variance of firm j at time t conditional on relevant information being released; $\sigma^2_{OVRNT:NoNews,jt}$ is the overnight return variance of firm j at time t conditional on no relevant information; $\sigma^2_{TRDNG:News,jt}$ is the trading day return variance of firm j at time t conditional on relevant information being released; $\sigma^2_{TRDNG:NoNews,jt}$ is the trading day return variance of firm j at time t conditional on no relevant information; and p represents the pseudo-probability of news and no-news days. Equation (1) is written as an approximation because overnight and trading day returns may be correlated, in other words, prices may not follow a random walk.

Assuming that the return volatility due to public information is independent of other sources of return volatility, equation (1) can be rewritten as a regression equation:

$$\sigma^2_{Interval,jt} = \alpha + \beta_{OVRNT:News} I_{OVRNT:News,jt} + \beta_{TRDNG:News} I_{TRDNG:News,jt} + \beta_{TRDNG:NoNews} I_{TRDNG:NoNews,jt} + \varepsilon_{jt} \quad (2)$$

where $\sigma^2_{Interval,jt}$ is the squared daily return, $I_{OVRNT:News,jt}$ is 1 if relevant information is made public overnight; $I_{TRDNG:News,jt}$ is 1 if relevant information is made public during the trading day; and $I_{TRDNG:NoNews,jt}$ is 1 if no relevant information is made public during the trading day. Because equation (2) pools the time-series and the cross-section together, in the analysis we include fixed effects for firms and time.

The overall variance contribution of news is a product of (i) the intensity of news arrival, and (ii) the impact of news upon arrival. The parsimonious model above allows us to estimate the impact of news upon arrival, controlling for other drivers of variance. The frequency of identified news days during trading hours and overnight provides us with a measure of the intensity of news arrival. Intuitively, holding the level of overall variance constant, an increase in either of these two components means that a larger fraction of variance is explained by the arrival of public news.

Table 3A and 3B respectively provide estimates of the coefficients from regression equation (2). We use different measures of variance as the dependent variable: the first column uses raw returns, the second column uses excess returns over the market return, and the third through sixth columns use idiosyncratic returns from a market model regression with various combinations of fixed effects. The last column focuses on complex news events. The results are robust to all these specifications with the proviso that identified news explains less raw return variance than idiosyncratic variance because identified news are focused on capturing firm specific, not market-wide, events. For the discussion that follows, we therefore focus on the idiosyncratic volatility estimates provided in the sixth column, which includes firm and date fixed effects, as it controls for variations in variance across firms and time.

First and foremost, the return variance impact of news is positive and large. The coefficient $\beta_{OVRNT:News} = 4.75$ can be interpreted as the incremental variance contribution of public information during closing hours, while $\left(\beta_{TRDNG:News} - \beta_{TRDNG:NoNews} = 6.53 - 2.49 = 4.04\right)$ represents the variance contribution of public information during trading hours.[23] To gauge the economic magnitudes of these identified news deltas, 4.75 and 4.04, note that the unconditional variances during overnight and trading hours are respectively 1.47 and 3.68 (Panel B, sixth column). The economic contribution of news to stock return variance is

[23] As pointed out above in Section II.A and shown in Table 2, Panels B and C, the volatility overnight conditional on identified news is higher than the volatility during trading hours on no-news days, thus reversing French and Roll's (1986) result comparing day and night volatility. Table 3A demonstrates the efficacy of this finding using the regression equation (2) with, for example, $\beta_{OVNT:news}=4.75>2.49=\beta_{TRDNG:NoNews}$. This result is statistically significant and holds across the various specifications (including those presented below in Table 5 below)

therefore large, with the relative increase of variance due to news more pronounced during overnight hours.

Second, as shown in Table 1 and repeated here in Table 3B, the fraction of overnight news days (15.32%) is marginally higher than those of trading hours (11.32%). Coupled with the incremental news contribution result discussed above, the contribution of news to overnight volatility is much greater compared to its contribution during trading hours. Specifically, 49.59% of overnight return volatility is explained by identified news compared to only 12.43% of the volatility during trading hours.[24]

Third, the final column parses out the identified news even further by focusing on complex news days. Specifically, in Panel A, the variance impact of complex news is two-three times higher, $\beta_{OVRNT:ComplexNews} = 11.35$ and $\left(\beta_{TRDNG:ComplexNews} - \beta_{TRDNG:NoNews} = 11.48 - 2.47 = 9.01\right)$. Panel B shows that the news deltas are magnitudes higher on these days. Even though complex news days cover only 27% (i.e., $^{4.19\%}/_{15.32\%}$) and 22% (i.e., $^{2.48\%}/_{11.32\%}$) of news days for overnight and trading hours, respectively, 32.4% and 6.1% of idiosyncratic variance is explained, i.e., 65.3% and 48.8% of the identified news days' contribution to variance.

We compare this analysis with the one that uses the *Ravenpack* data source. The results are reported in Table A2, panels A and B, of the Appendix. Again, the qualitative results remain unchanged, providing further confirmation of the two approaches to the extent they apply different textual methodologies. The evidence broadly suggests that public news have a more significant role than previously considered to volatility, even outside of a small set of meaningful news days (i.e., earning announcements).

Beyond the relative magnitude of news variance contribution during trading hours and overnight, the results can be further broken down by the category of firm-level public news using the classifications provided in Appendix A.[25] Table 4 reports these findings. The most

[24] Overnight news variance contribution is equal to overnight fraction of news days (15.32%) times overnight news delta (4.75) divided by unconditional residual returns squared (1.47). The same holds for all news variance contribution calculations.

[25] Note that the total fraction of identified news days is less than the sum of the fraction for each news category because some days include news on multiple categories. The implication of these multiple news days

striking result from Table 4 is that, irrespective of the news coming out during trading versus overnight hours, the relative importance of the specific category of news is preserved. In particular, the correlation across the 18 categories between trading and overnight variance contribution is 90.8%. In other words, even though the percentage contribution to variance is much higher overnight due to (i) a greater fraction of news, (ii) a higher news delta (impact), and (iii) lower overall volatility in overnight hours, the most important sources of news remain consistent. For example, the five same most important sources show up overnight and during trading hours (albeit in different order), respectively, *Financial*, *Forecasts*, *Mergers & Acquisitions*, *Earnings Factors*, and *Ratings* compared to *Financial*, *Ratings*, *Earnings Factors*, *Forecasts*, and *Mergers & Acquisitions*.

An additional interesting result is the effective zero correlation between an event being more likely and the event having greater impact. For overnight news, the most likely events in order are *Financial*, *Deal*, *Mergers & Acquisitions*, *Earnings Factors*, and *Forecast*, while the greatest impact ones are *Stocks*, *Ratings*, *Forecast*, *Financing* and *Financial*. For example, news about *Deals*, such as service and product deals, licensing, contracts and contract bids, partnerships, memorandum of understanding, pacts, joint ventures, collaborations, agreements, development partnerships, and technology implementation, are ranked second in frequency during both overnight and trading hours, yet only 15th and 13th out of 18, respectively, in terms of impact.

In addition to the cross-section of event types, there is also a wide cross-section of firm characteristics across the S&P500. Some of the non-valuation based characteristics include firm size, volume and firm age. A large literature documents in some form or another young and small stocks having greater volatility. The argument revolves around whether these firms are substantially more prone to either mispricing, or higher risk premia, or possessing more dynamic fundamentals due to their life cycle. (See, for example, Berk (1995), Chordia and Swaminathan (2000), Pastor and Veronesi (2003), Fama and French (2004), Brown and Kapadia (2007), Chun, Kim, Morck and Yeung (2008), and Fang and Peress (2009), among others.) Viewing these firm characteristics in light of the fraction of total overnight variance

---

is that the specific event-level news contribution variance will be an upper bound due to the assumption of full contribution by each event on these days.

attributable to news may provide insights on the drivers of return volatility and mispricing. While a detailed analysis is beyond the scope of this paper, we ask whether these cross-sectional characteristics are correlated with each firm's identified news contribution to volatility (albeit for S&P500 firms). We follow standard sorts and break the sample into large versus small firms, high versus low volume and old versus young firms. We also consider a double sort based on size and volume.

Table 5 reports return variance decomposition of news for the breakdown of the main firm characteristics described above. Some interesting stylized facts emerge. Perhaps not surprisingly, size is an important factor describing the relative importance of news. On the one hand, the incremental delta of news is considerably higher for small relative to large firms both during overnight as well as during trading hours (e.g., 8.43 versus 3.52 overnight and 8.93 versus 2.81 during trading hours). In other words, news matters more for smaller firms. On the other hand, there is a much greater likelihood of relevant news being recorded for larger firms during both overnight and trading hours (e.g., 18.46% versus 9.50% overnight and 14.07% versus 6.36% during trading hours). While these effects offset, the more dominant factor is the frequency of news, with total return variance being explained by public news being equal to 50.74% and 12.85% for large firms versus 43.13% and 11.74% for small firms, during overnight and trading hours, respectively.

Firm size is highly correlated with firm volume. To analyze the volume characteristic separately, we adjust for firm size. That is, we assign firms into volume bins conditional on their size bins. We confirm that the double sort indeed results in volume being independent of size by computing the cross-sectional correlation of the orthogonalized characteristics with size (e.g., the empirical correlation is 0.013 for volume). Consider overnight returns. In contrast to the size characteristic above, high volume firms have both higher incremental news deltas overnight (e.g., 5.19 versus 4.26), as well as higher frequency of identified news (20.86% versus 9.99%). This leads to more return variance explained by public news overnight (e.g., 56.96% versus 40.68%). Similar results hold for trading returns albeit lower in magnitude.

These results suggest that firm characteristics play an important role in trading, information revelation and return volatility. They also highlight that for subsets of stocks, arrival of

public information in the form of news articles is a major source of overall volatility. Firm age also presents interesting results. The delta impact of news is higher for young firms but the likelihood of an identified news event is lower. The combination leads to approximately equal contribution to volatility for young and old firms, respectively, 50.2% versus 49.2% for overnight returns and 11.9% versus 13.8% for trading day returns. However, the stylized fact that the delta impact of news is 5.94 and 5.11 for young firms overnight and during the trading day, respectively versus 3.70 and 3.28 for older firms supports dynamic models of firms (e.g., Pastor and Veronesi (2003) and Chun, Morck and Yeung (2008)).

As mentioned above, this analysis assumes that overnight and trading day returns are uncorrelated, so that there are no spillover effects. In other words, the directional impact of overnight news is fully incorporated in the overnight return. To address the possibility of return spillover from one period to the next, we extend the four possibilities in regression equation (2) into eight possible states by conditioning also on the previous period being either a no-news versus news period. The results are reported in Table 6.

Most important, panel A of Table 6 shows only small spillover from trading day to overnight. The regression coefficients for overnight volatility are similar irrespective of what happened during the earlier trading day. For example, focusing again on the regression based on residuals with firm and date fixed effects (last column of Table 6), the coefficients on identified news days are 4.67 conditioning on lagged news versus 5.15 conditioning on no lagged news. What matters is whether an identified news event occurs overnight or not (e.g., 5.15 versus a coefficient of 1.02 for no news in the third row). In contrast, the evidence supports a spillover from overnight to the trading day. The regression coefficients are also higher on trading days with identified news, but considerably so if identified news occurred the previous overnight (e.g., 9.50 versus 4.56 comparing rows 4 and 5 of column 6).

Panel B of Table 6 extends these findings further. Note that Panel B effectively breaks up identified news-driven volatility into three pieces: (i) the volatility level, (ii) likelihood of news, and (iii) the impact of news arrival. Focusing overnight and on the last column of the panel, the volatility on an identified news day is unconditionally much higher when no news come out during the trading day (i.e., 5.75 versus 2.49). The higher unconditional volatility

is presumably due to overnight news being "new" if no news preceded it during the trading day. The impact of this news relative to no news days is also higher (i.e., 5.15 versus 3.66 conditional on no news during the earlier trading hours versus news) for similar reasons. The likelihood of news is much higher (i.e., 10.70% versus 4.62%) conditional on no news because no news periods are themselves more likely. These three pieces together explain why identified news' contribution to overnight volatility is higher following no news compared to periods when news events have occurred during the day (e.g., 41.18% versus 6.79%).

V. Applications

A. R^2 and Noise Trading

Complementary to the analysis of variance ratios across different trading periods is the question of how much of the variation in stocks prices is due to fundamental information about the firm versus aggregate market. A key paper on the question of whether stock prices reflect fundamental information is Roll (1988) (see also French and Roll (1986) and Black (1986)). In that paper, Roll (1988) argues that once aggregate effects have been removed from a given stock returns, the finance paradigm would imply that the remaining return volatility would be idiosyncratic. As a proxy for this firm specific information, Roll (1988) uses news stories generated in the financial press. His argument is that, on days without news, idiosyncratic information is low, and the R^2s from aggregate level regressions should be much higher than on news days. Roll (1988) finds little discernible difference, leading to the well-known R^2 puzzle. Working off this result, a number of other papers reach similar conclusions with respect to prices and news, in particular, Cutler, Poterba and Summers (1989), and Mitchell and Mulherin (1994).

Here, we duplicate the analysis of Roll (1988) to help understand the relation between news and returns. Broadly, we document two key findings using our more precise identification of news, with one result contradicting Roll (1988) and one further deepening Roll's (1988) puzzle. In particular, we find that when news appear *and* are relevant (i.e., can be identified), news does matter. However, there are not enough identified news events (and

aggregate movements) to fully explain stock returns. Moreover, we provide some evidence in the cross-section of expected returns that supports a noise trading hypothesis.

The documented stylized fact in Table 2, that variances are higher on days in which we can identify important events and on days with complex news, supports a relation between prices and fundamentals. As a more formal analysis, we reproduce the aforementioned Roll (1988) analysis for our setting. We estimate a one-factor pricing model and a four-factor pricing model separately for each firm and for each day classification: all, no news, unidentified news, identified and identified complex news.[26] The R^2 are adjusted for the number of degrees of freedom.

Table 7 reports median R^2 across firms for the different news types. Consider the median calculations for the one-factor model. The R^2s are similar on no news and unidentified news days (i.e., 34.5% vs. 33.9%). The magnitude of the R^2s and the similarity of these numbers between no news and news days (albeit unidentified) are consistent with Roll's puzzling results. However, R^2s are much lower on identified news days, i.e., 17.7%. The difference in R^2 between identified news and no-news days is striking – the ratio of median R^2 between identified news and no-news days is 1.95, in sharp contrast to Roll's results. Similar to the results from Table 2 with respect to variance ratios, the results in Table 7 are even more pronounced on complex news days, with R^2s being lower, i.e., 10.0%. These results appear to be robust to the pricing model. For example, using the typical four-factor model (the market, book-to-market, size and momentum factors), the ratio of median R^2 between no-news and identified news days is only slightly lower (1.89 versus 1.95), and the R^2s between no-news and unidentified days is again similar.

Even though the drop in R^2s from no news days to identified news days is impressive, substantial unexplained variability in stock returns remains. That is, on days without news about the company on the Dow Jones wire, either identified or unidentified, the market (or four factor) regressions still only explain 34.5% (40.2%) of stock return variation. This finding suggests either a behavioral explanation or one based on trading-based private

[26] We impose a minimum of 40 observations to estimate the regressions.

information revelation.[27] We try to differentiate the behavioral from the private information explanation by repeating the R^2 analysis for trading hours and overnight. This analysis is novel to the literature.

As described above, a popular explanation for the large spread between variance ratios during trading hours and overnight is the revelation of information through trading. This explanation has been offered for the surprisingly low R^2s on no news days (and, in our paper, unidentified news days) of a regression of stock returns on multiple factors. To evaluate this explanation further, we run factor regressions using trading hour returns and overnight returns, conditional on various types. These results are also reported in Table 7.

The results strongly support the hypothesis that when important public information is identified, this information matters for stock prices. During closing hours, when less trading takes place, R^2s for identified news and complex news days are 10.6% and 6.6%, respectively, compared to 17.2% and 14.7% during trading hours. That is, by conditioning on overnight periods with relevant public information but, by construction, with little private information trading or noise trading, the explanatory power of aggregate factors drops.

On the other hand, the results also deepen the behavioralist view, that there is a large amount of unexplained stock price variability. During closing hours, when private information revelation through trading is unlikely to be a source for unexplained variability, conditioning on either no news or unidentified news, R^2s are only 39.5% or 37.3% respectively. Furthermore, these R^2s are not that much higher than the R^2s of 27.6% and 25.8% during trading hours. These results fine-tune and extend the challenge for rational pricing.

To this point, De Long, Shleifer, Summers and Waldmann (1990) develop a model with noise trading that links asset price volatility and higher expected returns to the level of noise

[27] An alternative explanation not pursued here is that idiosyncratic stock returns movements of firm I may be large on no news days if a related firm J (e.g., in the same sector) has an identified news event, such as mergers, financial forecasts, earnings, etc. In other words, identified news on a subset of economically linked firms may in effect have relevant news for the remaining firms.

trading. While their paper focuses on aggregate systematic risk, subsequent work also looks at the cross-section of stock returns (see, for example, Pontiff (1996), Daniel, Hirshleifer and Subrahmanyam (1998), Baker and Wurgler (2006) and Kumar and Lee (2006), among others).

Overnight, private information induced trading is unlikely to be the driver of variance given the low volume after hours. Thus, a natural interpretation of the fraction of overnight variance which is not news-driven is one of a proxy for mispricing. We run a pooled cross-sectional regression year-by-year of average risk-adjusted returns of S&P500 firms on each firm's news variance contribution. These regressions are performed using various combinations of overnight and trading day variance contributions with and without various controls (such as size, value and momentum).[28]

Table 8 reports the results. The first three columns demonstrate a significant negative relation between expected returns and identified news explained volatility derived from overnight returns. For example, interpreting the -0.875 in the first column, a one-standard deviation increase in the overnight news contribution to the variance of a stock (i.e., 7.03%) implies a 6.15% lower expected return on the stock in the following year. Columns 4-6 repeat the analysis for trading day returns. While the same result for the cross-section of expected returns holds, the economic magnitude is much smaller, with a one standard deviation shock leading to a 1% lower expected return. Indeed, when overnight and trading day variance contributions are both included (in the final column of Table 8), the overnight effect remains negative and statistically significant while the trading day effect becomes statistically indistinguishable from zero.

In other words, if one takes the view that unexplained overnight idiosyncratic volatility is driven by "noise", firms subject to greater noise trading have higher expected returns, an implication consistent with the noise trading hypothesis. For the trading day, the weaker relation between the cross-section of expected returns and unexplained volatility may be

[28] The controls represent dummy variables if the firm is in the top two quintiles of size, book-to-market or momentum.

due to idiosyncratic volatility being driven by private information-based trading, not just noise.

B. Information Production and the Resolution of Uncertainty

Morck, Yeung and Yu (2000) document higher market model R^2s for individual firm stock returns in emerging countries, as well as a general decline in market model R^2s in the U.S. from 1926 to the late 1990s. They argue and show support for the theory that better property rights (in developed countries and over time) lead to information production by arbitrageurs. This firm specific information is incorporated into stock prices and increases idiosyncratic volatility (see also Campbell, Lettau, Malkiel and Xu (2001)). This type of argument is formalized by Veldkamp (2006a, 2006b). In her model, with high fixed costs of information production, it is only worthwhile for arbitrageurs to produce information that impacts many firms, leading to more correlated stock returns in the cross-section and higher market model R^2s. As these costs decrease (or alternatively the benefits rise), more firm-specific information gets produced and idiosyncratic risk takes over.

In an update to their paper, Morck, Yeung and Wu (2013) document an interesting result. Over the last 15 or so years, the downward trend in market model R^2s for individual firm stock returns in the U.S. has reversed, and, financial crisis aside, has been drifting upward. For our sample of S&P500 firms, Table 9 documents year-by-year measures of the average level of idiosyncratic firm return variance and market variance, as well as the news variance contribution.

Even though our sample of firms (i.e., S&P500) differ from those of Morck, Yeung and Wu (2013), columns 2-4 of Table 9 confirm the result that both the level of idiosyncratic variance and the ratio of average idiosyncratic variance to overall market variance decrease from 2000 to 2015. For example, the coefficient of these variables with a linear time trend are -0.419 and -0.154, respectively (last row of the table). In the Veldkamp (2006a, 2006b) framework, this finding would correspond to either increasing costs or decreasing benefits to information production.

As additional evidence, consider the case of overnight returns which better isolates public from private information-based volatility. To the extent an increase in variance contribution from public news reduces the benefit of information production, Veldkamp's model implies a corresponding decrease in the relative level of idiosyncratic variance.[29] Column 5 of Table 9 also shows a downward trend in average idiosyncratic variance during overnight hours with a time trend coefficient of -0.112 while, simultaneously, the contribution of identified news to return volatility has an upward time trend with a significant coefficient of 0.020.

To further investigate this result, consider trading day returns which in theory should incorporate private-information based trading. If we ignore noise trading and remove public information-based (i.e., identified news) variance, all that remains is private-information induced variance. Column 9 of Table 9 documents the time-series of the idiosyncratic variance level during trading hours arising from private information. Consistent with lower private information production, idiosyncratic variance trends downward with a time trend coefficient of -0.311.

If the value of private information relative to public information is indeed lower, then in Veldkamp's (2006a) framework two conditions need to be present. The first is that, as Veldkamp (2006a) points out, an independent volatility shock (i.e., one that does not resolve uncertainty) will increase the demand for information production. Arbitrageurs, however, will be less willing to produce this information if there is a surge in public information after volatility shocks, due either to firms disseminating this information or the media producing news. The second condition is that the public information should resolve uncertainty, i.e., lead to future lower volatility. As pointed out by Campbell, Lettau, Malkiel and Xu (2001), improved information about firm specific cash flows resolves uncertainty and decreases idiosyncratic return variance (though clearly increasing price volatility on the information release). This result would be interesting as it goes against standard autoregressive conditional heteroskedasticity (ARCH) models of volatility. That is, higher volatility today (i.e., arising from the public information shock) does not necessarily lead to

[29]Other explanations for lower information production have been suggested. For example, Doidge, Karolyi and Stulz (2017) document a downward trend in publicly listed companies, thus potentially reducing the benefit of private information because information can be applied to fewer companies.

high volatility next period. Table 10 documents stylized facts with respect to both of these implications.

With respect to the first implication, we construct a measure of the shock to idiosyncratic variance (week *t* relative to prior 4 weeks) on non-identified news days. In Veldkamp's (2006a) framework, high values of these measures should lead to greater private information production. The value of such information, however, will be diminished if large idiosyncratic volatility shocks also lead to higher future intensity of public information, i.e., identified news.

Table 10, Panel A reports a regression of the change in future identified news days on a spike in variance on unidentified news days.[30] The coefficients are positive and statistically significant. That is, spikes in variance (of unknown origin) lead to more public information. Consider column 2 of Panel A. The coefficient of 0.37, along with a standard deviation of 6.45, means that a one standard deviation increase in the volatility spike leads to 2.36% more identified news days (relative to the unconditional value of 21.6%), representing a 10.9% increase in news intensity. Ceteris paribus, the substantive increase in public information reduces the value of private information and thus is one possible explanation for the lower idiosyncratic variance.

With respect to the second implication, Panel B in Table 10 compares the time-series pattern of stock return variances (relative to the variance on all days) on identified news days versus unidentified news days with similarly high variance shocks (e.g., 2.76 on identified news days versus 2.63). On days leading up to the identified news, the variance ratio is higher than normal, going from 1.39, 1.39, 1.45, 1.49 to 2.16 on day t-5 to t-1. Interestingly, and key to the above story, the variance ratio drops to 1.4, 1.27, 1.23, 1.21 and 1.20 over the following five days after the public information. In other words, while identified news produces significant volatility on the event day, this public information is for the most part fully revealed and subsequent volatility falls which, again, is consistent with a reduction in the value of private information-based trading.

[30] The change in future identified news count is measured relative to the previous 20 days, either measured over the next day or averaged over the following week. The variance spike is measured as a variance jump of 1.5 or 2.0 times the variance over the past 20 days.

In contrast, shocks on no-news days or unidentified news days are more consistent with the partial revelation of private information. While the variance ratio is high prior to the price shock at date t, i.e., 1.79, 1.82, 1.82, 1.86 to 1.72 from t-5 to t-1, the variance ratio is similarly high after the shock, i.e., 1.90, 1.85, 1.82, 1.83 and 1.80 from date t+1 to t+5. In other words, there seems to be a fundamental difference in the time-series behavior of variance ratios around stock price shocks depending on whether the information is publicly identified or not.

Putting aside our potential explanation for the findings of Morck, Yeung and Yu (2000, 2013), the different time-series patterns of variances conditional on volatility shocks due to identified versus non-identified news are interesting their own right. The primary models of volatility, namely those related to autoregressive conditional heteroskedasticity (ARCH), may need to be adjusted depending on the source of the volatility shock.

VI. **Conclusion**

Innovations in textual analysis allow researchers to better identify the relevance and content of news. Using a supervised learning methodology to identify firm-level events from the Dow Jones newswire, we provide an empirical methodology that allows us to isolate the portion of return variance due solely to the arrival of these events. The key takeaway is that, when relevant news can be identified, stock prices are closely linked to this news. Examples of results include variance ratios of returns on identified news days that are more than double those on no news and unidentified news days, and even more so overnight; incremental explained variance from public information around 50% overnight and 10% during trading hours; and model R^2s that are no longer the same on news versus no news days, but now are 17% versus 35%.

The paper, however, documents variance ratio patterns, market model R^2s, and relative variance contributions during overnight and trading hours, that in some way deepen the excess volatility puzzle described and analyzed in the literature. The information identifier methodology described in this paper may be useful for a deeper analysis of the relation between stock prices and information. For example, there is a large literature that looks at stock return predictability and reversals/continuation of returns depending on under-reaction

or over-reaction to news (see, for example, Hirshleifer (2000), Chan (2003), Vega (2006), Gutierrez and Kelley (2008), Tetlock, Tsaar-Tsechansky, and Macskassy (2008), and Tetlock (2010)). This paper allows the researcher to segment this news into categories likely to lead to under- or over-reaction.

Moreover, a vast literature in behavioral finance argues that economic agents, one by one, and even in the aggregate, cannot digest the full economic impact of news quickly. Given this database of identified events, it is possible to measure and investigate "complexity" and its effect on the speed of information processing by the market. For example, "complexity" can be broken down into whether more than one economic event occurs at a given point in time, how news (even similar news) gets accumulated through time, and cross-firm effects of news. We hope to explore some of these ideas in future research.

References

Admati, A. and P. Pfleiderer, 1988, A theory of Intraday Patterns: Volume and Price Variability, *Review of Financial Studies* 1, 3–40.

Aharony, J., and I. Swary, 1980, Quarterly Dividend and Earnings Announcements and Stockholders' Returns: An Empirical Analysis, *Journal of Finance* 35, 1–12.

Antweiler, W., and M. Z. Frank, 2005, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance* 59, 1259–1293.

Asquith, P., and D. W. Mullins, 1986, Equity Issues and Offering Dilution, *Journal of Financial Economics* 15, 61–89.

Baker, M. and Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance*, *61*(4), pp.1645-1680.

Ball, R., and P. Brown, 1968, An Empirical Evaluation of Accounting Income Numbers, *Journal of Accounting Research* 6, 159–178.

Barber, B., and Odean, T., 2008. All that glitters: the effect of attention and news on the buying behavior of individual and institutional investors. Review of Financial Studies 21, 785–818.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. "Twitter mood predicts the stock market." *Journal of computational science*2, no. 1 (2011): 1-8.

DellaVigna, S., and Pollet, J., 2006. Investor inattention, firm reaction, and friday earnings announcements. Journal of Finance 64, 709–749.

Barclay, M.J., Litzenberger, R.H., and J.B. Warner, 1990, Private information, trading volume, and stock-return variances, *Review of Financial Studies*, 3, 233-254.

Barclay, M., and T. Hendeshott, 2003, Price Discovery and Trading After Hours, *the Review of Financial Studies*, 16(4), 1041-1073.

Berk, J. B., 1995, A Critique of Size-Related Anomalies, *Review of Financial Studies* 8(2), 275-286.

Berry, T. D., and K. M. Howe, 1994, Public Information Arrival, *Journal of Finance* 49, 1331–1346.

Black, F., 1986, Noise, *Journal of Finance* 41, 529–543.

Bradley, Daniel, Jonathan Clarke, Suzanne Lee, and Chayawat Ornthanalai, 2014, "Are analysts' recommendations informative? Intraday evidence on the impact of time stamp delays." *Journal of Finance* 69, no. 2: 645-673.

Brown, Gregory, and Nishad Kapadia, 2007, "Firm-specific risk and equity market development." *Journal of Financial Economics* 84.2: 358-388.

Campbell, J. Y., 1991, A Variance Decomposition for Stock Returns, *Economic Journal* 101, 157–179.

Campbell, J.Y., Lettau, M., Malkiel, B.G. and Xu, Y., 2001. Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. *Journal of Finance*, *56*(1), pp.1-43.

Dougal, Casey, Joseph Engelberg, Diego Garcia, and Christopher A. Parsons. "Journalists and the stock market." *The Review of Financial Studies* 25, no. 3 (2012): 639-679.

Chan, W. S., 2003, Stock Price Reaction To News And No-News: Drift And Reversal After Headlines, *Journal of Financial Economics* 70, 223–260.

Chen, Hailiang, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. "Wisdom of crowds: The value of stock opinions transmitted through social media." *Review of Financial Studies* 27, no. 5 (2014): 1367-1403.

Chordia, T., R. Roll and A. Subrahmanyam, 2011, Recent Trends in Trading Activity and Market Quality, *Journal of Financial Economics*, 101/2, 243263.

Chordia, Tarun, and Bhaskaran Swaminathan, 2000, "Trading volume and cross-autocorrelations in stock returns." *Journal of Finance* 55.2: 913-935.

Chun, Hyunbae, Jung-Wook Kim, Randall Morck, and Bernard Yeung, 2008, "Creative destruction and firm-specific performance heterogeneity." *Journal of Financial Economics* 89, no. 1: 109-135.

Cohen, L. and Lou, D., 2012. Complicated firms. *Journal of financial economics*, *104*(2), pp.383-400.

Cutler, D. M., J. M. Poterba, and L. H. Summers, 1989, What Moves Stock Prices?, *Journal of Portfolio Management* 15, 4–12.

Cooper, M. J., M. T. Cliff, and H. Gulen, 2008, Return differences between trading and non-trading hours: Like night and day, Available at SSRN 1004081.

Daniel, K., D. Hirshleifer, and A. Subrahmanyam, 1998, Investor Psychology and Security Market under- and Overreactions, *Journal of Finance* 53, 1839–1885.

Das, S. R., and M. Y. Chen, 2007, Yahoo! For Amazon: Sentiment Extraction from Small Talk on The Web, *Management Science* 53, 1375–1388.

Davis, A. K., J. Piger, and L. M. Sedor, 2012, Beyond the Numbers: An Analysis of Optimistic And Pessimistic Language In Earnings Press Releases, *Contemporary Accounting Research* 29, 845–868.

De Long, J.b., A. Shleifer, L. H. Summers, and R. J. Waldmann, 1990, Noise Trader Risk in Financial Markets, *Journal of Political Economy* 98(4), 703-738.

Demers, E., and C. Vega, 2010, Soft Information In Earnings Announcements: News Or Noise?,Working Paper, INSEAD.

Doidge, C., Karolyi, G.A. and Stulz, R.M., 2017. The US listing gap. *Journal of Financial Economics*, *123*(3), pp.464-487.

Engelberg, J. E. , 2008, Costly Information Processing: Evidence From Earnings Announcements, Working Paper, University of North Carolina.

Engle, R. F, M. Hansen, and A. Lunde, 2011, And Now, The Rest of The News: Volatility and Firm Specific News Arrival, Working Paper.

Fama, E. F., L. Fisher, M. C. Jensen, and R. Roll, 1969, The Adjustment of Stock Prices to New Information, *International Economic Review* 10, 1–21.

Fama, Eugene F., and Kenneth R. French, 2004, "New lists: Fundamentals and survival rates." *Journal of financial Economics*73.2: 229-269.

Fang L., and J. Peress, 2009, Media Coverage and the Cross-Section of Stock Returns, *Journal of Finance* 64(5), 2023-2052.

Feldman, R., S. Govindaraj, J. Livnat, and B. Segal, 2010, Managements Tone Change, Post Earnings Announcement Drift and Accruals, *Review of Accounting Studies* 15, 915–953.

Feldman, R., and J. Sanger, 2006, The Text Mining Handbook, Cambridge University Press.

Fleming, J., C. Kirby and B. Ostdiek, 2006, Stochastic Volatility, Trading Volume, and the Daily Flow of Information, *Journal of Business*, 79/3, 1551-1590.

Foster, F., and S. Viswanathan, 1990, A theory of intraday variations in volumes, variances and trading costs in securities markets, *Review of Financial Studies*, 3, 593-624.

Francis, J., D. Pagach, and J. Stephan, 1992, The stock market response to earnings announcements released during trading versus nontrading periods, *Journal of Accounting Research*, 165-184.

French, K. R., and R. Roll, 1986, Stock Return Variances: The Arrival of Information and Reaction of Traders, *Journal of Financial Economics* 17, 5–26

Garcia, Diego. "Sentiment during recessions." *Journal of Finance* 68, no. 3 (2013): 1267-1300.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy. *Text as data*. No. w23276. National Bureau of Economic Research, 2017.

Glosten, L. R., and P. R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14(1), 71-100.

Greene, J. T., and S. G. Watts, 1996, Price discovery on the NYSE and the NASDAQ: The case of overnight and daytime news releases. Financial Management 25, 19-42.

Griffin, J. M., N. H. Hirschey, and P. J. Kelly, 2011, How Important Is The Financial Media In Global Markets?, *Review of Financial Studies* 24, 3941–3992.

Grob-Klubmann, A., and N. Hautsch, 2011, When Machines Read the News: Using Automated Text Analytics to Quantify High Frequency News-Implied Market Reactions, *Journal of Empirical Finance* 18, 321–340.

Grossman, S. J., and J. E. Stiglitz, 1980, On the impossibility of informationally efficient markets *The American Economic Review*, 393-408.

Gutierrez, R. C., and E. K. Kelley, 2008, The Long-Lasting Momentum in Weekly Returns, *Journal of Finance* 63, 415–447.

Hanley, K. W., and G. Hoberg, 2012, Litigation Risk, Strategic Disclosure and The Underpricing Of Initial Public Offerings, *Journal of Financial Economics* 103, 235–254.

Hansen, Stephen, Michael McMahon, and Andrea Prat. "Transparency and deliberation within the FOMC: a computational linguistics approach." *The Quarterly Journal of Economics* (2014).

Heston, Steven L., and Nitish Ranjan Sinha. "News versus sentiment: Comparing textual processing approaches for predicting stock returns." *Robert H. Smith School Research Paper* (2014).

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4), 2013.

Hirshleifer, D., 2001, Investor Psychology and Asset Pricing, *Journal of Finance*, 56, 1533– 1597.

Hoberg, Gerard, and Gordon Phillips. "Product market synergies and competition in mergers and acquisitions: A text-based analysis." *Review of Financial Studies* 23, no. 10 (2010): 3773-3811.

Hoberg, Gerard, and Gordon Phillips. "Text-based network industries and endogenous product differentiation." *Journal of Political Economy* 124, no. 5 (2016): 1423-1465.

Hong, H. and J. Stein, 2003, Differences of Opinion, ShortSales Constraints, and Market Crashes, *Review of Financial Studies*, 16 (2): 487-525.

Ito, T., R. Lyons and M. Melvin, 1998, Is There Private Information in the FX Market? The Tokyo Experiment, *Journal of Finance*, 53: 1111-1130.

Jegadeesh, Narasimhan, and Di Wu. "Word power: A new approach for content analysis." *Journal of Financial Economics* 110, no. 3 (2013): 712-729.

Jiang, C., T. Likitapiwat, and T. McInish, 2012, Information Content of Earnings Announcements: Evidence from After-Hours Trading, *Journal of Financial and Quantitative Analysis*, Volume 47, pp 1303-1330.

Jones, C.M., Kaul, G., and M.L. Lipson, 1994. Information, trading, and volatility, *Journal of Financial Economics*, 36, 127-154.

Kelly, M. A., and S. P. Clark, 2011, Returns in trading versus non-trading hours: The difference is day and night, *Journal of Asset Management* 12(2), 132-145.

Kogan, S., Routledge, B. R., Sagi, J. S., and N. A. Smith, 2011, Information Content of Public Firm Disclosures and the Sarbanes-Oxley Act, Working Paper.

Kumar, A. and Lee, C., 2006. Retail investor sentiment and return comovements. *Journal of Finance*, *61*(5), pp.2451-2486.

Kyle, A. S., 1985, Continuous auctions and insider trading, *Econometrica*, 1315-1335.

Li, Feng. "Annual report readability, current earnings, and earnings persistence." *Journal of Accounting and economics* 45, no. 2-3 (2008): 221-247.

Li, Feng. "The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach." *Journal of Accounting Research* 48, no. 5 (2010): 1049-1102.

Lou, D., C. Polk, and S. Skouras, 2018, "A Tug of War: Overnight Versus Intraday Expected Returns," working paper.

Loughran, T., and B. McDonald, 2011, When Is a Liability Not A Liability? Textual Analysis, Dictionaries, And 10-Ks, *Journal of Finance* 66, 35–65.

Loughran, Tim, and Bill McDonald. "Measuring readability in financial disclosures." *Journal of Finance* 69, no. 4 (2014): 1643-1671.

Madhavan, A., Richardson, M., and M. Roomans, 1997, Why do security prices fluctuate? A transaction-level analysis of NYSE stocks, *Review of Financial Studies*, 10, 1035-1064.

Mandelker, G., 1974, Risk and Return: The Case Of Merging Firms, *Journal of Financial Economics* 1, 303–335.

Manela, Asaf, and Alan Moreira. "News implied volatility and disaster concerns." *Journal of Financial Economics* 123, no. 1 (2017): 137-162.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Vol. 1, no. 1. Cambridge: Cambridge university press, 2008.

Milgrom, P., and N. Stokey, 1982, Information, trade and common knowledge, *Journal of Economic Theory* 26(1), 17-27.

Mitchell, M. L., and J. H. Mulherin, 1994, The Impact of Public Information on the Stock Market, *Journal of Finance* 49, 923–950.

Mitkov, Ruslan, 1999, Anaphora resolution: the state of the art, School of Langauge and Eropena Studies, University of Wolverhamton.

Morck, Randall, Bernard Yeung, and Wayne Yu. 2000, The information content of stock markets: why do emerging markets have synchronous stock price movements? *Journal of Financial Economics* 58.1: 215-260.

Morck, Randall, Bernard Yeung, and Wayne Yu. 2013, R2 and the Economy, *Annual Reviews of Financial Economics*, Vol. 5: 145-166.

Neuhierl, A., A. Scherbina, and B. Schlusche, 2013, Market Reaction to Corporate Press Releases, *Journal of Financial and Quantitative Analysis*, forthcoming.

Pástor, Ľuboš, and Veronesi Pietro, 2003, Stock valuation and learning about profitability, *Journal of Finance* 58.5: 1749-1789.

Pontiff, Jeffrey, 1996, Costly arbitrage: Evidence from closed-end funds, *The Quarterly Journal of Economics* 111.4: 1135-1151.

Price, S. McKay, James S. Doran, David R. Peterson, and Barbara A. Bliss. "Earnings conference calls and stock returns: The incremental informativeness of textual tone." *Journal of Banking & Finance* 36, no. 4 (2012): 992-1011.

Roll, R., 1984, Orange Juice and Weather, *American Economic Review* 74, 5, 861–880.

Roll, R., 1988, R2, *Journal of Finance* 43, 541–566.

Shiller, R. J., 1981, The Use of Volatility Measures in Assessing Market Efficiency, *The Journal of Finance* 36(2), 291-304.

Shleifer, A., 2000, Inefficient Markets: An Introduction to Behavioral Finance, Oxford University Press, Oxford.

Tauchen, G. E. and M. Pitts, 1983, The price variability volume relationship on speculative markets, *Econometrica*, 51, 485-505.

Tetlock, P. C., 2007, Giving Content to Investor Sentiment: The Role of Media in The Stock Market, *Journal of Finance* 62, 1139–1168.

Tetlock, P. C., 2010, Does Public Financial News Resolve Asymmetric Information? *Review of Financial Studies* 23, 3520–3557.

Tetlock, Paul C. "All the news that's fit to reprint: Do investors react to stale information?", *Review of Financial Studies* 24, no. 5 (2011): 1481-1512.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy, 2008, More Than Words: Quantifying Language To Measure Firms' Fundamentals, *Journal of Finance* 63, 1437–1467.

Vega, C., 2006, Stock Price Reaction to Public and Private Information, *Journal of Financial Economics* 82, 103–133.

Veldkamp, L.L., 2006a. Media frenzies in markets for financial information. *The American economic review*, *96*(3), pp.577-601.

Veldkamp, L.L., 2006b. Information markets and the comovement of asset prices. *The Review of Economic Studies*, *73*(3), pp.823-845.

Wisniewski, Tomasz Piotr, and Brendan Lambe. "The role of media in the credit crunch: The case of the banking sector." *Journal of Economic Behavior & Organization* 85 (2013): 163-175.

Appendix A: Topic Identification

Business Trends: Customer Traffic Increase, Customer Traffic Decrease, Consumer Spending Positive, Consumer Spending Negative, Market Share Increase, Market Share Decrease

Capital Returns: Stock Buy Back, Dividends

CSR/Brand: Scandal, Corruption, Social Responsibility, Environmental Responsibility, Company Image, Brand Image, Product Image, Credibility, Award, Sponsorship, Affiliations

Deals: Service Deals, Product Deals, Licensing, Contract Bid, Alliance, Partnerships, MOU, Pacts, Joint Ventures, Collaborations, Contracts, Agreements, Development Partnership, Tech Implementation

Earnings Factors: Tailwinds, Headwinds, Challenges & Pressures, Emerging Market Positive, Emerging Market Negative, Commodity Production, Manufacturing, Competition, Capacity Change, Currency Pressure, Currency Tailwind, Weather Positive, Weather Negative

Employment: CEO Change, Executive Change, Board Change, Executive Compensation, Employment Issues, Strikes, Workforce Increase, Workforce Decrease

Facility: Facilities Opening, Facilities Closing

Financial: Financial Results, Financial Results Beat, Financial Results Miss, Financial, Bankruptcy, Margin Expansion, Margin Pressure, Inventory Increase, Inventory Decrease, Restructuring

Financing: Equity Offering, Debt Financing, Private Equity Financing

Forecast: Guidance Change Positive, Guidance Change Negative, Forecast Negative, Forecast Positive, Non - Company Forecast, Company Growth, Company Growth Slowdown, Innovation, Category Expansion

General: Activists Actions, Company Update, Investment

Investment

Legal: Investigation, Indictments, Arrest Charges, Suspension Ban, Fraud, Money Laundering, Bribery, Settlement, Judgement, Lawsuit, Legal Issues, Regulatory issues

Mergers & Acquisitions: Merger & Acquisition, Acquisition of Startups, Merger of Equals, Asset Acquisition, Asset Sale, Synergy

Product: Product Recall, Product Trial Results, Product Approval, New Product Launch, Product Issues, Product Update, Product Technology, Product Functions, Product Features

Ratings: Analyst Rating, Price Target, Credit - Debt Rating, Rating Agency List

Stocks: Spin Off - Split Off, IPO

Stock Holdings: Fund Position, Inside Sell-Purchase, Index Change

Tables

Table 1: Summary Statistics

Panel A: DAY

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Total	1,952,175	896	3.4	319	106
No News	1,112,341	895	NA	NA	NA
Unid News	417,889	888	1.9	313	70
Iden News	421,945	888	4.9	321	119
Complex News	124,824	873	9.2	330	129

Panel B: OVRNT

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Total	1,946,458	895	2.8	327	114
No News	1,338,742	895	NA	NA	NA
Unid News	308,620	887	1.7	316	75
Iden News	299,096	886	3.8	333	132
Complex News	81,803	862	6.9	345	146

Panel C: TRDNG

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Total	1,946,458	895	2.3	307	94
No News	1,432,131	894	NA	NA	NA
Unid News	293,353	886	1.7	308	62
Iden News	220,974	886	3.1	307	117
Complex News	48,439	841	5.6	317	134

The table reports summary statics on the number of tickerdate observations, the number of unique tickers, the average number of articles, words per article and relevant words per article. No News days are days on which no news appeared, Unidentified News days are days on which news appeared but did not contain any identified event (see Appendix A for a list of events), Complex News days are identified news days on which more than two different identified events (or sub-events) appeared. Panel A includes tickerdate definitions based on a close-close window ("DAY"), Panel B includes tickerdate definitions based on a close-open window ("OVRNT"), and Panel C includes tickerdate definitions based on an open-close window ("TRDNG").

Table 2: Event Frequency Across Return Ranks and Variances

Panel A: DAY

	Return Rank			Stock SD and Variance		
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio
Total	1.0%	-0.6%	0.1%	2.31	896	1.22**
No News	-7.9%	0.6%	3.3%	2.04	890	1.00
Unid News	0.5%	-0.3%	0.1%	2.17	849	1.16**
Iden News	24.9%	-4.2%	-8.3%	2.96	839	2.17**
Complex News	63.3%	-11.3%	-20.4%	4.16	703	4.23**

Panel B: OVRNT

	Return Rank			Stock SD and Variance		
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio
Total	1.0%	-0.7%	0.2%	1.20	895	1.49**
No News	-7.0%	1.0%	2.5%	0.92	893	1.00
Unid News	-1.9%	-0.3%	1.3%	1.05	831	1.26**
Iden News	39.6%	-8.4%	-11.4%	2.09	818	4.76**
Complex News	98.9%	-21.3%	-28.2%	3.10	628	10.11**

Panel C: TRDNG

	Return Rank			Stock SD and Variance		
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio
Total	1.0%	-0.6%	0.1%	2.02	895	1.08**
No News	-4.3%	0.2%	2.0%	1.90	893	1.00
Unid News	7.5%	-1.5%	-2.3%	2.07	819	1.25**
Iden News	26.6%	-4.1%	-9.2%	2.51	779	1.78**
Complex News	57.5%	-9.1%	-19.6%	3.01	457	2.42**

We assign daily returns into percentiles separately for each stock and year: bottom/top 10% (i.e., extreme 20% of returns), moderate 40% of return moves, and the smallest 40% return moves. For each of the first three columns, we compare the observed intensity of different day types to the intensity predicted under the null that these distributions are independent. The next three columns report the median standard deviation (per day type), the number of unique tickers, and the median variance ratio (across tickers), i.e., the median ratio (across firms) of squared return deviations on each day type divided by the squared deviations on no news days. For a description of day types, see Table 1. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 3: News Variance Contribution

Panel A: Variance Regressions

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2	Eps^2
			News = IdenNews				News = High Inten News
$I_{OVRNT,News}$	4.544	4.300	4.236	4.693	4.324	4.754	11.351
	[0.145]***	[0.140]***	[0.138]***	[0.071]***	[0.069]***	[0.071]***	[0.126]***
$I_{TRDNG,News}$	7.133	6.126	5.979	6.514	5.987	6.529	11.480
	[0.139]***	[0.132]***	[0.130]***	[0.081]***	[0.078]***	[0.081]***	[0.161]***
$I_{TRDNG,NoNews}$	3.329	2.566	2.466	2.476	2.480	2.485	2.466
	[0.030]***	[0.028]***	[0.028]***	[0.038]***	[0.038]***	[0.038]***	[0.035]***
Constant	1.259	0.869	0.818	0.000	0.000	0.000	0.000
	[0.020]***	[0.020]***	[0.020]***	NA	NA	NA	NA
Observations	3,892,864	3,892,864	3,892,864	3,892,864	3,892,864	3,892,864	3,892,864
R^2	0.003	0.003	0.002	0.003	0.003	0.003	0.004
Fixed Effects	None	None	None	Firm	Date	Firm & Date	Firm & Date
$\beta_{OVRNT,News} = \beta_{TRDNG,NoNews}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Panel B: Variance Firm-Level News Component

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2	Eps^2
Fixed Effects	None	None	None	Firm	Date	Firm & Date	Firm & Date
OVRNT (unconditional mean)	1.96	1.53	1.47	1.47	1.47	1.47	1.47
TRDNG (unconditional mean)	5.02	3.84	3.68	3.68	3.68	3.68	3.68
OVRNT, frac of News days	15.32%	15.32%	15.32%	15.32%	15.32%	15.32%	4.19%
TRDNG, frac of News days	11.32%	11.32%	11.32%	11.32%	11.32%	11.32%	2.48%
OVRNT News Δ	4.54	4.30	4.24	4.69	4.32	4.75	11.35
TRDNG News Δ	3.80	3.56	3.51	4.04	3.51	4.04	9.01
OVRNT News Var Contribution	35.57%	43.07%	44.18%	48.95%	45.10%	49.59%	32.38%
TRDNG News Var Contribution	8.58%	10.50%	10.80%	12.41%	10.78%	12.43%	6.07%

Panel A of the table reports panel regressions in which the dependent variable are various squared firm and time window returns: $R^2_{DAY,jt} = \alpha + \beta_{OVRNT:News} I_{OVRNT:News,jt} + \beta_{TRDNG:News} I_{TRDNG:News,jt} + \beta_{TRDNG:NoNews} I_{TRDNG:NoNews,jt} + \epsilon_{jt}$. In columns 1 these are raw returns, in column 2 these are excess returns (relative to the market), and in columns 3-6 these are residual returns from a one-factor market model. The independent variables include a dummy for close-open identified news days ($I_{OVRNT:News}$), a dummy for open-close news days ($I_{TRDNG:News}$), and a dummy for open-close no-identified news days ($I_{TRDNG:NoNews}$). Columns 4-6 include firm, date, or both fixed-effects. Panel B of the table reports the unconditional means of the squared returns during non-trading ("$OVRNT$") and trading ("$TRDNG$") hours, the fraction of identified news days during the two time windows, the Δ that is due to identified news during the two time windows, and the overall contribution of identified news to variance. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 4: News Variance Contribution by Event Type

	OVRNT			TRDNG		
	frac of News days	News Δ	News Var Contribution	frac of News days	News Δ	News Var Contribution
Identified News	15.32%	4.75	49.59%	11.32%	4.04	12.43%
BusinessTrend	0.61%	7.96	3.32%	0.42%	4.40	0.51%
CSRBrand	0.70%	3.48	1.66%	0.47%	1.49	0.19%
CapitalReturns	1.19%	6.79	5.51%	0.82%	3.65	0.81%
Deal	3.66%	3.75	9.34%	2.22%	1.87	1.13%
EarningsFactors	2.47%	9.29	15.61%	1.67%	7.75	3.51%
Employment	1.69%	6.50	7.46%	1.22%	3.89	1.29%
Facility	0.83%	7.97	4.48%	0.54%	4.31	0.63%
Financial	4.76%	9.40	30.46%	3.06%	8.21	6.83%
Financing	0.54%	12.22	4.45%	0.36%	12.97	1.26%
Forecast	2.15%	12.49	18.31%	1.56%	7.26	3.07%
General	0.11%	4.16	0.32%	0.07%	1.37	0.03%
Investment	0.49%	9.32	3.08%	0.29%	5.90	0.46%
Legal	1.81%	5.18	6.37%	1.21%	4.88	1.60%
MergerAcquisition	2.85%	8.48	16.43%	1.73%	5.06	2.38%
Product	1.83%	3.53	4.39%	1.13%	1.04	0.32%
Rating	1.14%	13.35	10.32%	1.43%	12.21	4.75%
Stock	0.09%	22.05	1.38%	0.05%	10.92	0.16%
StockHoldings	1.04%	2.77	1.96%	0.53%	1.72	0.25%

The table reports panel regressions in which the dependent variable are residual returns from a one-factor market model: $R^2_{DAY,jt} = \alpha + \beta_{OVRNT:News} I_{OVRNT:News,jt} + \beta_{TRDNG:News} I_{TRDNG:News,jt} + \beta_{TRDNG:NoNews} I_{TRDNG:NoNews,jt} + \epsilon_{jt}$. We run these regressions for each event type separately. That is, we assign the dummies $News, jt$ to be equal to one if the event type is observed during the window (OVRNT or TRDNG). We include firm and date fixed-effects. The table reports the fraction of identified news days during the two time windows, the Δ that is due to identified news during the two time windows, and the overall contribution of identified news to variance.

Table 5: News Variance Contribution and Firm Characteristics

	Size		Volume		AGE	
	Coeff.	SdErr.	Coeff.	SdErr.	Coeff.	SdErr.
I_{High}	0.276	0.082	0.614	0.067	1.204	0.129
$I_{OVRNT,News}$	8.433	0.153	4.255	0.117	5.942	0.105
$I_{High,OVRNT,News}$	-4.908	0.174	0.938	0.147	-2.240	0.146
$I_{TRDNG,News}$	12.144	0.184	5.683	0.135	8.148	0.123
$I_{High,TRDNG,News}$	-7.295	0.206	1.445	0.168	-2.910	0.167
$I_{TRDNG,NoNews}$	3.215	0.066	2.011	0.051	3.035	0.055
$I_{High,TRDNG,NoNews}$	-1.171	0.081	1.019	0.075	-1.074	0.078
	Small	Large	Low	High	Young	Old
OVRNT News Δ	8.433	3.524	4.255	5.194	5.942	3.703
TRDNG News Δ	8.929	2.805	3.672	4.099	5.113	3.277
OVRNT (uncon. mean)	1.86	1.28	1.045	1.902	1.76	1.21
TRDNG (uncon. mean)	4.84	3.07	2.899	4.482	4.44	2.99
OVRNT, % News days	9.50%	18.46%	9.99%	20.86%	14.83%	16.12%
TRDNG, % News days	6.36%	14.07%	7.31%	15.48%	10.28%	12.58%
OVRNT News Var Cont.	43.13%	50.74%	40.68%	56.96%	50.17%	49.22%
TRDNG News Var Contr.	11.74%	12.85%	9.26%	14.15%	11.85%	13.79%

The table reports panel regressions with the dependent variable being the one-factor market model return residuals squared and the independent variables includes dummies for highlow characteristic (changing across columns) interacted with dummies for return window with identified news and those without: $R^2_{DAY,jt} = \alpha + \beta_{OVRNT:News} I_{OVRNT:News,jt} + \beta_{TRDNG:News} I_{TRDNG:News,jt} + \beta_{TRDNG:NoNews} I_{TRDNG:NoNews,jt} + \beta_{High} * I_{High,jt} + \beta_{High:OVRNT:News} I_{High:OVRNT:News,jt} + \beta_{High:TRDNG:News} I_{High:TRDNG:News,jt} + \beta_{High:TRDNG:NoNews} I_{High:TRDNG:NoNews,jt} + \epsilon_{jt}$. All regressions include firm and date fixed effects. The bottom part of Panel A reports the unconditional means of the squared return residuals overnight ("$OVRNT$") and during trading hours ("$TRDNG$"), the fraction of identified news days during the two time windows, the Δ that is due to identified news during the two time windows, and the overall contribution of identified news to variance. Size dummy is equal to 1 if the firm size quantile assignment is equal to 5, and 0 otherwise (recall that the majority of firms are in size quantile 5 since they are S&P500 firms). Volume dummy is equal to one if the firm volume is greater than its annual average volume, Coverage dummy is equal to 1 if the firm's coverage is greater than the median coverage for the same size group that year, BM dummy is equal to 1 if the firm book-to-market quantile assignment is greater than 3 and 0 if it is less than 3, MOM dummy is equal to 1 if the momentum quantile assignment is greater than 3 and 0 if it is less than 3, and AGE dummy is equal to 1 if the firm age is greater than the median age of firms that year. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 6: News Variance Contribution by Lagged Interval Type

Panel A: Variance Regressions

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2
$I_{OVRNT,News,LagNews}$	4.088	3.655	3.523	4.651	3.514	4.672
	[0.287]***	[0.270]***	[0.261]***	[0.124]***	[0.119]***	[0.124]***
$I_{OVRNT,News,LagNoNews}$	4.798	4.630	4.597	5.022	4.754	5.147
	[0.165]***	[0.161]***	[0.160]***	[0.082]***	[0.081]***	[0.082]***
$I_{OVRNT,NoNews,News}$	0.508	0.457	0.457	0.896	0.565	1.015
	[0.185]***	[0.185]**	[0.186]**	[0.101]***	[0.100]***	[0.101]***
$I_{TRDNG,News,LagNews}$	9.745	8.612	8.424	9.473	8.426	9.503
	[0.263]***	[0.247]***	[0.242]***	[0.118]***	[0.114]***	[0.118]***
$I_{TRDNG,News,LagNoNews}$	5.080	4.168	4.053	4.505	4.096	4.561
	[0.131]***	[0.125]***	[0.124]***	[0.104]***	[0.103]***	[0.104]***
$I_{TRDNG,NoNews,News}$	4.158	3.356	3.258	3.690	3.417	3.816
	[0.063]***	[0.055]***	[0.053]***	[0.084]***	[0.082]***	[0.084]***
$I_{TRDNG,NoNews,LagNews}$	3.266	2.503	2.403	2.404	2.409	2.410
	[0.028]***	[0.026]***	[0.026]***	[0.040]***	[0.040]***	[0.039]***
Constant	1.219	0.833	0.782	0.000	0.000	0.000
	[0.015]***	[0.014]***	[0.014]***			
Observations	3,892,864	3,892,864	3,892,864	3,892,864	3,892,864	3,892,864
R^2	0.003	0.003	0.003	0.003	0.003	0.003
Fixed Effects	None	None	None	Firm	Date	Firm & Date

Panel B: Variance Firm-Level News Component

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2
Fixed Effects	None	None	None	Firm	Date	Firm & Date
OVRNT, LagNews (unconditional mean)	3.19	2.60	2.49	2.49	2.49	2.49
OVRNT, LagNoNews (unconditional mean)	7.23	5.93	5.75	5.75	5.75	5.75
TRDNG LagNews (unconditional mean)	4.62	3.46	3.31	3.31	3.31	3.31
TRDNG, LagNoNews (unconditional mean)	3.58	3.20	3.07	3.76	2.95	3.66
OVRNT, LagNews (frac news)	4.62%	4.62%	4.62%	4.62%	4.62%	4.62%
OVRNT, LagNoNews (frac news)	10.70%	10.70%	10.70%	10.70%	10.70%	10.70%
TRDNG LagNews (frac news)	5.08%	5.08%	5.08%	5.08%	5.08%	5.08%
TRDNG, LagNoNews (frac news)	6.24%	6.24%	6.24%	6.24%	6.24%	6.24%
OVRNT News Δ , LagNews	3.58	3.20	3.07	3.76	2.95	3.66
OVRNT News Δ , LagNoNews	4.80	4.63	4.60	5.02	4.75	5.15
TRDNG News Δ , LagNews	5.59	5.26	5.17	5.78	5.01	5.69
TRDNG News Δ , LagNoNews	1.81	1.67	1.65	2.10	1.69	2.15
OVRNT News Var Contribution, LagNews	5.19%	5.69%	5.69%	6.97%	5.47%	6.79%
OVRNT News Var Contribution, LagNoNews	28.54%	35.58%	36.78%	40.18%	38.04%	41.18%
TRDNG News Var Contribution, LagNews	3.93%	4.50%	4.56%	5.11%	4.42%	5.02%
TRDNG News Var Contribution, LagNoNews	2.45%	3.00%	3.11%	3.96%	3.18%	4.06%

Panel A of the table reports panel regressions in which the dependent variable are various squared firm and time window returns. The returns are measured over intraday intervals (either OVRNT or TRDN) and are pooled. The independent variables include dummies that interact the interval type (OVRNT or TRDNG), whether identified news were released during that interval (News or NoNews), and whether the lagged interval contained identified news (LagNews or LagNoNews). In columns 1 these are raw returns, in column 2 these are excess returns (relative to the market), and in columns 3-6 these are residual returns from a one-factor market model. Panel B of the table reports the unconditional means of the squared returns, the fraction of identified news days across all classifications, the Δ that is due to identified news during the two time windows, and the overall contribution of identified news to variance. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 7: R^2s – Firm-level Regressions

	DAY					OVRNT			TRDG	
	N	Median R^2	Ratio	Median R^2	Ratio	N	Median R^2	Ratio	N	Median R^2	Ratio
	Single Factor Regressions			Four Factor Regressions			Single Factor Regressions				
Total	896	27.9%	1.24**	32.6%	1.24**	895	24.2%	1.63**	895	25.5%	1.08**
No News	886	34.5%	1.00	40.2%	1.00	889	39.5%	1.00	891	27.6%	1.00
Unid News	812	33.9%	1.02**	39.1%	1.03**	779	37.3%	1.06**	756	25.8%	1.07**
Iden News	800	17.7%	1.95**	21.3%	1.89**	770	10.6%	3.73**	707	17.2%	1.60**
Complex News	569	10.0%	3.46**	14.3%	2.81**	479	6.6%	6.02**	298	14.7%	1.87**

The table reports results from firm level return regressions, across a number of different specifications. In all regressions, the dependent variable is time t firm return. Columns 1-5 report the results for close-close ("DAY"), columns 6-8 report the results for overnight ("OVRNT"), and columns 9-11 report the results for trading hours ("TRDNG"). We use 1 and 4 factor models. The values reported in the table are the median R^2s, across stocks, and the ratio of the median R^2 relative to the R^2 on no-news days, and the number of observations. For a description of day types, see Table 1. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 8: News Variance Contribution and Firm Returns

	Excess Returns Year $(t+1)$						
OVRNT NewsVar Cont	-0.875	-0.648	-0.697				-0.993
	[0.160]***	[0.161]***	[0.162]***				[0.278]***
TRDG NewsVar Cont				-3.197	-3.108	-2.334	0.768
				[0.870]***	[0.861]***	[0.886]***	[1.453]
OVRNT Res^2		0.006					
		[0.002]***					
TRDG Res^2					0.006		
					[0.002]***		
Size Q			-1.671			-1.793	
			[0.806]**			[0.804]**	
BM Q			0.255			0.215	
			[0.298]			[0.297]	
Mom Q			-0.839			-0.848	
			[0.465]*			[0.465]*	
Constant	5.048	3.702	14.117	4.681	3.553	14.493	5.022
	[0.482]***	[0.564]***	[4.283]***	[0.461]***	[0.504]***	[4.267]***	[0.483]***
Observations	6,944	6,944	6,646	6,946	6,946	6,647	6,940
R^2	0.002	0.012	0.004	0.001	0.014	0.003	0.002

The table reports cross sectional regressions of firms' next year average daily excess return, measured using beta-adjusted residual returns, on current year measures of OVRNT(TRDNG) news variance contribution, OVRN(TRDNG) average squared residuals returns, market cap quintile, book-market quintile, and momentum quintile. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 9: News Variance Contribution by Year

Year	DAY			OVRNT		TRDNG		
	Idios Var	Market Var	Ratio	Idios Var	News Contrib	Idios Var	News Contrib	News Unxplained Idios Var
2000	10.87	2.19	4.97	3.03	25.40%	8.48	2.81%	8.24
2001	8.38	1.94	4.32	2.61	32.12%	6.76	9.72%	6.11
2002	9.28	2.77	3.35	2.63	42.25%	7.08	12.89%	6.17
2003	3.61	1.09	3.31	1.10	36.79%	2.83	9.59%	2.56
2004	2.44	0.49	4.98	0.79	43.17%	1.86	8.48%	1.70
2005	2.42	0.42	5.73	0.89	60.89%	1.74	14.79%	1.49
2006	2.26	0.40	5.66	0.91	42.74%	1.77	11.85%	1.56
2007	2.51	1.00	2.51	0.93	43.61%	2.07	10.71%	1.85
2008	13.17	6.76	1.95	4.19	51.34%	10.81	17.23%	8.95
2009	7.91	2.81	2.81	2.17	48.55%	6.30	13.93%	5.42
2010	2.17	1.27	1.71	0.70	52.77%	1.86	7.77%	1.71
2011	2.37	2.09	1.13	0.71	60.00%	1.81	11.29%	1.60
2012	2.17	0.64	3.37	0.67	67.19%	1.56	12.34%	1.36
2013	1.69	0.50	3.40	0.73	52.39%	1.23	11.32%	1.09
2014	1.63	0.50	3.25	0.65	60.95%	1.20	7.58%	1.11
2015	2.41	0.94	2.57	0.87	59.72%	1.77	6.69%	1.65
Time trend coeff	-.419**	-.055	-.154**	-.112**	.020***	-.333**	.001	-0.311**

The table reports various statistics for each year of our sample. Columns 2-4 report the average beta-adjusted squared residual returns (i.e., idiosyncratic variance), market squared returns, and the ratio of the two, respectively, over daily intervals (close-close). Columns 5-6 report the average beta-adjusted squared residual returns and news variance contribution overnight ("OVRNT"). Column 7-8 report the same statistics for trading hours ("TRDNG"). Column 9 reports the level of idiosyncratic variance not explained by news during trading hours. The last row of the table reports the coefficient obtained by regressing each of these series on a linear time variable. ***(**) denote p-values lower than 1% (5%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table 10: Media Frenzy

Panel A

Conditioning	No IdenNews at time t and volatility jump of > 1.5		No IdenNews at time t and volatility jump of > 2.0	
Dependent Variable	$t+1$ IdeNews	$t+1$ to $t+5$ IdeNews	$t+1$ IdeNews	$t+1$ to $t+5$ IdeNews
($\times$100)				
time t volatility spike	0.366	0.225	0.332	0.203
	[0.035]***	[0.020]***	[0.040]***	[0.021]***
Constant	-1.102	0.112	-0.830	0.289
	[0.146]***	[0.093]	[0.187]***	[0.111]***
Observations	287,065	286,799	212,750	212,564
R^2	0.021	0.020	0.030	0.027
SD of the independent variable	6.452	6.452	7.314	7.314

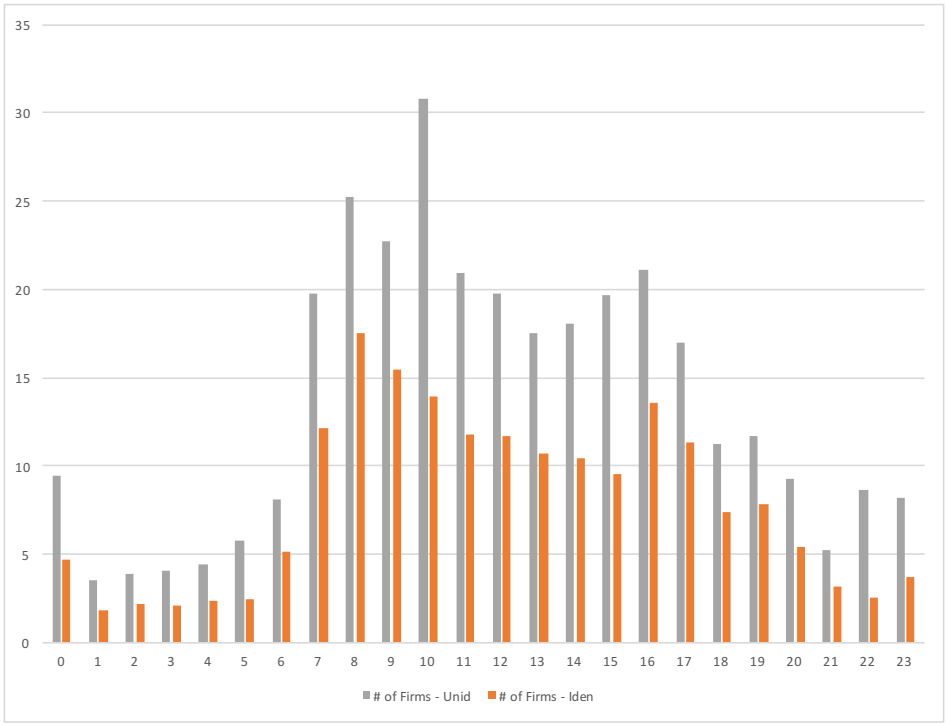
Panel B

Day	$t-5$	$t-4$	$t-3$	$t-2$	$t-1$	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
Non-IdenNews at time t with abnormal volatility (.5)	1.787	1.818	1.815	1.857	1.720	2.633	1.901	1.854	1.820	1.828	1.798
IdenNews at time t	1.387	1.390	1.447	1.490	2.163	2.761	1.421	1.268	1.227	1.209	1.195

Panel A of the table regresses abnormal identified news intensity (average daily count of identified news relative to lagged 20 days) either at the daily or weekly (5 day) horizon on lagged volatility spikes and day type. The first two columns condition on the firm having no lagged identified news days and a variance jump greater than 1.5 (relative to pervious 20 days) and the next two columns condition on the firm having no lagged identified news days and a variance jump greater than 2.0 (relative to pervious 20 days). Below the regression results we report the standard deviation of the independent variable under both specifications. Panel B of the table reports median of variance ratios relative to time t where it is either defined by firms having no identified news but a variance spike greater than 0.5 or by having identified news. Variance ratios are calculated using beta-adjusted residual returns. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

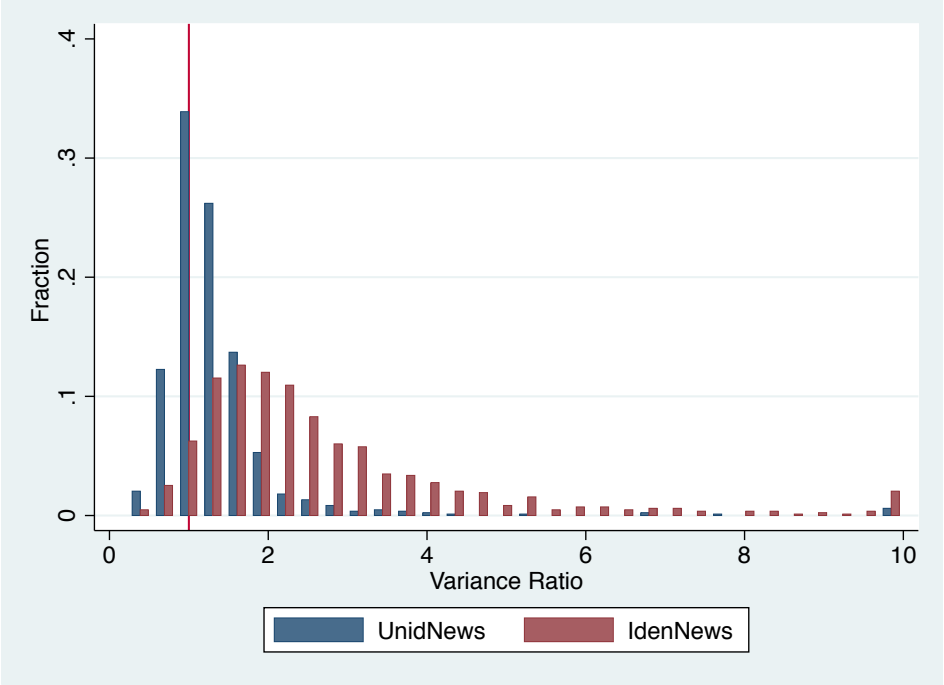
Figures

Figure 1: Hourly News Distribution



The figure plots the distribution of Unidentified and Identified news intensity over the day and over the week. For each hour of the day, we compute the average number of firms with identified news or with unidentified news.

Figure 2: Distribution of Variance Ratios by Day Type



The figure depicts the distribution of variance ratios, calculated within stocks, of unidentified and identified news days relative to no news days. Ratios are winsorized at 10. For a description of day types, see Table 1.

A Appendix

Table A.1: Event types – Summary Statistics

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Acquisition	22,270	724	8.6	302	76
Analyst Rec	12,411	680	8.5	335	66
Deals	30,101	718	6.8	315	93
Employment	21,489	741	6.3	283	87
Financial	69,205	783	7.6	309	71
Legal	10,764	581	8.6	291	71
Partnerships	10,047	587	7.3	371	110
Product	25,181	652	7.1	366	108

	Stock Return (daily)	Market Ret (daily)	SIZE	BM	MOM
Acquisition	10.4bp	-1.7bp	4.81	2.91	2.81
Analyst Rec	-21.7bp	0.7bp	4.75	2.87	2.77
Deals	9.2bp	-1.3bp	4.81	2.92	2.84
Employment	-5.3bp	-1.3bp	4.74	3.00	2.70
Financial	-0.4bp	0.1bp	4.73	2.89	2.83
Legal	-3.7bp	1.1bp	4.85	2.82	2.67
Partnerships	8.7bp	0.4bp	4.84	2.66	2.88
Product	6.7bp	-1.1bp	4.81	2.70	2.82

The table groups day/ticker observations by appearance of each of the event types (acquisitions, analyst recommendations, deals, employment, financial, partnerships, and products) and reports, in the top panel, the number of observations, the total number of ticker, the average number of articles, the average number of words per article, and the average number of relevant words per article. The bottom panel uses the same classification and reports the average daily returns, the average CRSP Value Weighted Market return, and the returns on the size, book-to-market, and momentum factors.

Table A.2: Event Frequency Across Return Ranks and Variances – Ravenpack

Panel A: DAY

	Return Rank			Stock SD and Variance			
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio	Obs
Total	1.0%	-0.6%	0.1%	2.62	745	1.26**	1,162,221
No News	-11.2%	1.1%	4.5%	2.18	616	1.00	252,897
Unid News	-3.9%	0.3%	1.7%	2.44	710	1.15**	708,857
Iden News	33.5%	-5.6%	-11.2%	3.60	662	2.52**	200,467

Panel B: TRDNG

	Return Rank			Stock SD and Variance			
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio	Obs
Total	1.0%	-0.5%	0.1%	2.30	745	1.14**	1,162,221
No News	-8.0%	0.4%	3.6%	2.04	701	1.00	452,261
Unid News	4.5%	-0.8%	-1.4%	2.37	706	1.30**	621,713
Iden News	21.9%	-3.4%	-7.6%	2.70	574	1.63**	88,247

Panel C: OVRNT

	Return Rank			Stock SD and Variance			
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio	Obs
Total	1.0%	-0.7%	0.2%	1.33	745	1.57**	1,162,221
No News	-11.2%	2.0%	3.6%	0.96	691	1.00	405,381
Unid News	-6.0%	0.7%	2.3%	1.11	708	1.21**	610,571
Iden News	63.8%	-14.1%	-17.8%	2.58	638	6.31**	146,269

The first three columns of the tables report the difference between the observed distribution of observations and that predicted under independence. We assign daily returns into percentiles separately for each stock and year: bottom/top 10% (i.e., extreme 20% of returns), moderate 40% of return moves, and the smallest 40% return moves. For each of these columns, we compare the observed intensity of different day types to the intensity predicted under the null that these distributions are independent. The next three columns report the median standard deviation (per day type), the number of unique tickers, and the median variance ratio (across tickers), i.e., the median ratio (across firms) of squared return deviations on each day type divided by the squared deviations on no news days. For a description of day types, see Table 1. **(*) denote p-values lower than 5% (10%) obtained from a non-parametric test of the null that the median variance ratio is equal to one.

Table A.3: News Variance Contribution – RavenPack

Panel A: Variance Regressions

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2
$I_{OVRNT,News}$	6.476	6.051	5.952	6.418	5.941	6.343
	[0.243]***	[0.236]***	[0.232]***	[0.107]***	[0.103]***	[0.107]***
$I_{TRDNG,News}$	8.354	7.068	6.952	7.514	6.950	7.472
	[0.231]***	[0.221]***	[0.221]***	[0.133]***	[0.129]***	[0.132]***
$I_{TRDNG,NoNews}$	4.620	3.581	3.448	3.465	3.446	3.458
	[0.044]***	[0.040]***	[0.039]***	[0.051]***	[0.051]***	[0.051]***
Constant	1.612	1.228	1.149	0.000	0.000	0.000
	[0.030]***	[0.028]***	[0.028]***	NA	NA	NA
Observations	2,324,442	2,323,498	2,323,498	2,323,498	2,323,498	2,323,498
R^2	0.004	0.003	0.003	0.003	0.003	0.003
Fixed Effects	None	None	None	Firm	Date	Firm & Date

Panel B: Variance Firm-Level News Component

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2
Fixed Effects	None	None	None	Firm	Date	Firm & Date
OVRNT (unconditional mean)	2.43	1.99	1.90	1.90	1.90	1.90
TRDNG (unconditional mean)	6.51	5.07	4.86	4.86	4.86	4.86
OVRNT, frac of News days	12.59%	12.59%	12.59%	12.59%	12.59%	12.59%
TRDNG, frac of News days	7.59%	7.59%	7.59%	7.59%	7.59%	7.59%
OVRNT News Δ	6.48	6.05	5.95	6.42	5.94	6.34
TRDNG News Δ	3.73	3.49	3.50	4.05	3.50	4.01
OVRNT News Var Contribution	33.58%	38.27%	39.45%	42.54%	39.38%	42.04%
TRDNG News Var Contribution	4.35%	5.22%	5.47%	6.32%	5.47%	6.27%

Panel A of the table reports panel regressions in which the dependent variable are various squared firm and time window returns: $R^2_{DAY,jt} = \alpha + \beta_{OVRNT:News} I_{OVRNT:News,jt} + \beta_{TRDNG:News} I_{TRDNG:News,jt} + \beta_{TRDNG:NoNews} I_{TRDNG:NoNews,jt} + \epsilon_{jt}$. In columns 1 these are raw returns, in column 2 these are excess returns, and in columns 3-6 these are residual returns from a one-factor market model. The independent variables include a dummy for close-open identified news days ($I_{OVRNT:News}$), a dummy for open-close news days ($I_{TRDNG:News}$), and a dummy for open-close no-identified news days ($I_{TRDNG:NoNews}$). Columns 4-6 include firm, date, and firmdate fixed-effecs. Panel B of the table reports the unconditional means of the squared returns during non-trading ("$OVRNT$") and trading ("$TRDNG$") hours, the fraction of identified news days during the two time windows, the Δ that is due to identified news during the two time windows, and the overvall contribution of identified news to variance.