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? Systematic factors aside, prior literature concludes that the revelation of private information through trading, and not public news, is the primary driver. We revisit the question by utilizing new textual analysis tools that allow us to better-identify fundamental information in news. We find that such fundamental firm-level information is an important source for stock price volatility, accounting for 20%-40% of overnight volatility (compared to 5%-6% during trading hours). Moreover, we find that the percentages of news-explained variance varies across firm characteristics and industries.

*Corresponding author: Shimon Kogan, Arison Business School, IDC Herzliya, Tel: +972 (9) 966-2741, email: *skogan@idc.ac.il*. We would like to thank 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, 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."

[a] Arison School of Business, IDC Herzliya.
[b] School of Business Administration, The Hebrew University.
[c] Arison School of Business, IDC Herzliya and the University of Texas at Austin.
[d] Stern School of Business, New York University and NBER.

1. 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 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]

As an illustration, French and Roll (1986) compare 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 argument is that private information can affect volatility only during trading hours as it is gradually revealed through trading. French and Roll (1986) conclude that the evidence strongly supports private-information rational trading models as the main driver of return volatility. Using more complete data and additional real-world experiments, this conclusion has generally been confirmed by, among others, Barclay, Lizenberger and Warner (1990), Ito, Lyons and Melvin (1998), Barclay and Hendershott (2003), Madhavan, Richardson and Roomans (1997), and Chordia, Roll and Subrahmanyam (2011).

An alternative view of this 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

[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), Barclay and Hendershott (2003), Fleming, Kirby and Ostidek (2006), Cliff, Cooper and Gulen (2008), Kelly and Clark (2011), and Lou, Polk and Skouras (2015), among others.

news..."; and Hong and Stein (2003) write "Roll (1984, 1988) and French and Roll (1986) demonstrate in various ways that it is hard to explain asset price movements with tangible public information" (p.487).

Our paper provides a contrasting view to both these literatures. 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. This is done within the standard framework – stock return variances during trading and overnight. Using textual analysis, we identify *relevant* public information tied to specific firm events from news stories. We then re-evaluate existing findings by identifying informationally relevant non-trading and trading periods, thus controlling for private information induced volatility.[3]

Common to much of this literature, the proxy for public information has been news articles.[4] A problem with this proxy is its potentially low power. Common news sources for companies, such as those in the Wall Street Journal stories and Dow Jones News Service, *et cetera*, release many stories 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, that is a massive computational problem since there are hundreds of thousands, possibly millions, of news stories to work through. Fortunately, advances in the area of 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) an information extraction platform (Feldman, Rosenfeld, Bar-Haim and Fresko (2011), denote Feldman *et al*. (2011), called *The Stock Sonar* or *TSS*), which we make publicly available, and (ii) a machine-learning method that is an industry standard, *Ravenpack*. The paper focuses on the first approach, as

[3]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.

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

we are able to make the data for it publicly available, while replicating the main findings with the second approach. [5]

Using these two approaches, we match each news article, which itself is time-stamped and linked with stock ticker(s), to a series of "identified" value-relevant events, or are deemed "unidentified".[6] 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 degrees of intensity (to be defined precisely later on). Textual analysis allows us to identify which news is fundamental and this identification is key to our analysis. 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 news days, the volatility of stock prices is more than double that of other days, consistent with the idea that the intensity and importance of information arrival is not the same across these days. This is economically important since over two thirds of all major-media outlet news is *unidentified*.

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. This is an important distinction because, like French and Roll (1986), examining overnight separately helps control for volatility induced via private information-driven trading. In particular, we find that the variance ratio overnight of identified (high intensity) news to no news is 2.71 (5.55), a magnitude higher than that ratio during trading hours, namely 1.59 (2.11).

Consistent with the findings of French and Roll (1986) we find that unconditional median daily return volatility during trading hours is 2.30%, 73% higher than that overnight, namely, 1.33%. In contrast, the median volatility is 41% higher during the trading day

[5] The *TSS* dataset is available in an online appendix: https://www.dropbox.com/s/oagtni23qcqmaqb/Appendix-Part%20I%2020150222.pdf?dl=0.

[6] The *TSS* methodology, as well as *Ravenpack*, are discussed briefly in the body of the paper and in more detail in an online appendix.

relative to overnight (2.89% versus 2.05%) for identified news, and only 7% higher during intense news days (2.91% versus 2.72%). These findings provide a contrast to conclusions reached by French and Roll (1986) and others who unconditionally document considerable more volatility during trading hours.

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. This provides 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, high intensity news, and specific events) and compare the results from this decomposition both cross-sectionally, across industries and over time. 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 show that the variance explained by public information is around 20%-40% overnight and around 6% during trading hours.[7] These variance contributions are higher for firms that are larger, have higher trading volume, and higher coverage (most of which is explained by the greater quantity of news, as opposed to the impact of news). We also examine the average fraction of variance across industries explained by news. We show that there is a strong relation between the overall level of idiosyncratic variance within the industry and the fraction of that variance that is explained by news. Put differently, at least some of the difference in the level of idiosyncratic variance across industries can be explained by differences in the impact and intensity of news across these industries. In the time-series, we show that there is a large public information component during the 2008 crisis period. This component, however, is not due to an increased number of relevant news days but instead greater return variability on these days.

This paper is organized as follows. Section 2 describes the data employed throughout the study. Of special interest, we describe the textual analysis methodologies for identifying

[7] The exact fraction of news explained variance depends on the text extraction platform.

relevant news and lay out important stylized facts. Sections 3 and 4 provide the main results of the paper, showing a strong relationship between prices and news, once the news is appropriately identified. In addition, some existing anomalies are deepened once we separate out the identification of news. Section 5 concludes.

2. Data Description

A. Textual Analysis

With the large increase in the amount of daily news content on companies over the past decade, it should be no surprise that the finance literature has turned to textual analysis as a way to understand how information both arrives to the marketplace and relates to stock prices. Early work centered on document-level sentiment classification of news articles by employing pre-defined sentiment lexicons.[8] The earliest paper in finance that explores textual analysis is Antweiler and Frank (2005), who employ language algorithms to analyze internet stock message boards posted on "Yahoo Finance". Much of the finance literature, however, 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. A plethora of papers post Tetlock (2007) apply a similar methodology to measure the positive versus negative tone of news across a wide variety of finance and accounting applications.[9] Loughran and McDonald (2011), in particular, is interesting because they refine IV-4 to more finance-centric definitions of positive and negative words.[10]

[8] See, for example, Lavrenko, Schmill, Lawrie, Ogilvie, Jensen, and Allan (2000), Das and Chen (2007) and Devitt and Ahmad (2007), among others. Feldman and Sanger (2006) provide an overview.

[9] See, for example, Davis, Piger, and Sedor (2006), Engelberg (2008), Tetlock, Saar-Tsechansky and Macskassy (2008), Kothari, Li and Short (2009), Demers and Vega (2010), Feldman, Govindaraj, Livnat and Segal (2010), and Loughran and McDonald (2011), among others.

[10] More recently, an alternative approach to textual analysis in finance and accounting has been offered by Li (2010), Hanley and Hoberg (2011), Grob-Klubmann and Hautsch (2011) and Kogan, Routledge, Sagi and

The focus of this paper is quite different. In particular, we are not interested in identifying the sentiment of the news *per se*. Instead, 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*. We use two quite different methodologies to identify these events.

The first approach is a rule-based information extraction platform, described in Feldman *et al*. (2011). Feldman et al.'s (2011) platform, *The Stock Sonar* (*TSS*), was developed specifically to extract topics from financial news and it is available on commercial platforms like Dow Jones. TSS extracts event instances and sentiment out of the text based on a set of predefined rules. The initial list of events were chosen to match commercial providers such as *CapitalIQ* but were augmented by events likely to impact stock prices. This process led to a total of 14 event categories. The events fall into one of the following categories: *Analyst Recommendations*, *Financial*, *Financial Pattern*, *Acquisition*, *Deals*, *Employment*, *Product*, *Partnerships*, *Inside Purchase*, *Facilities*, *Legal*, *Award*, *Stock Price Change* and *Stock Price Change Pattern*. The methodology is described in some detail in the online appendix. Of potential interest to researchers, the online appendix also includes links to the ticker-event-date dataset used in this paper (full day, open to close, and close to open).

The second methodology comes from *RavenPack, Inc*., which represents an industry standard for asset management firms in terms of news analytics. In particular, *Ravenpack* uses machine learning algorithms to process text from Dow Jones newswire into a machine readable content in order to identify a company's news in terms of a "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.

While both methodologies apply fundamentally different approaches, the data used in this paper is derived from the same news articles and press releases that appear in the Dow

---

Smith (2011). These authors employ machine learning-based applications to decipher the tone and therefore the sentiment of news articles

Jones newswire (such as the Wall Street Journal). 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 *TSS* and the results are more conservative, for the remainder of the paper, we document results using the *TSS* identification of specific events. That said, for two of the key tables, Appendix B duplicate the findings using *RavenPack*. Any differences between the results are discussed in that context.

B. Data Set Construction

As mentioned above, 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, 2009. For computational reasons, and in order 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 at some time or another 791 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 total daily returns from CRSP.

In order 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 date *t* articles those that were released between 15:31 on date *t-1* and 15:30 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. Articles released overnight (weekends and holidays) are matched with the next available trading day. *TSS* methodology processes each article separately and generates an output file in which each article/stock/day is represented as an observation.

For each of the aforementioned observations, *TSS* 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). 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 *TSS* produces a list of relevant events connected to this company and to this particular piece of news. It is possible that multiple events may be connected to a given story. In our analysis, we ignore the *Stock Price Change* and *Stock Price Change Pattern* categories as these categories do not, on their own, represent fundamental news events. We also ignore *Award*, *Facilities*, and *Inside Purchase*, since these categories do not contain a sufficient number of observations.[11] We also merged *Financial* and *Financial Pattern* and are therefore left with eight main categories.

To be more precise, 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 categories:

1. *No news* – observations without news coverage.
2. *Unidentified news* – observations for which none of the news coverage is identified by one of the eight categories.[12]
3. *Identified news* – observations for which at least some of the news coverage is identified as being at least one of the above categories.

Moreover, conditional on being classified as *identified news*, we provide a further breakdown of identified news, with a subset of these days being defined as *high-intensity news* days, defined as *identified news* days with more than two event types (either categories or subcategories).[13]

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

[11] Including the Award, Facilities, and Inside Purchase categories does not alter the results.

[12] 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. Even under this classification, *RavenPack* identifies more "relevant" news stories than does TSS.

[13] Across the 14 event categories, there are 56 subcategories. For example, consider the *Analyst Recommendation* category. It contains nine subcategories, including *analyst expectation*, *analyst opinion*, *analyst rating*, *analyst recommendation*, *credit - debt rating*, *fundamental analysis*, *price target*, etc.

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).

It is, of course, possible that news lag the occurrence of the event mentioned in them. 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.

Table 1 provides an overview of the data. The first column in panel A reports the number of observations under each of the day classifications documented by *TSS*. Most days have no news coverage, i.e., 705,430 of 1,245,709 stock/day observations contain no news reported on the Dow Jones Newswire. As a comparison, the last column of Table A.1 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 252,897 days have no coverage. However, of some importance, the vast majority of the days with news coverage in *Ravenpack*, 708,857 of 909,324 have a relevance score less than 100% and are considered unidentified. Similarly, for the TSS methodology, 380,420 of 540,279 days do not have an identified topic news event. Most of the *TSS* identified news days contain only a single-identified event (i.e., 122,666 of 159,829) although these days may include several subcategories under the event.

Table 1 also observes that identified news days contain a larger number of articles compared with unidentified news days (6.1 vs. 2.6 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 TSS) is much larger on identified news days (81 vs. 49). Finally, of the 159,829 relevant news days, only 37,151 are high intensity news.

Panels B and C of Table 1 report similar statistics but now broken down between trading hours and overnight. The most striking result is the similarity between the two panels -- for the most part the news coverage is similar during periods when the market is open versus the market is closed. For example, the ratio of news days - unidentified, identified and high intensity - to total number of days - is respectively 25.6%, 8.6% and 2.3% during trading

hours versus 18.7%, 6.5%, and 1.8% overnight.[14] While this finding may have something to do with when news is reported, as opposed to when it takes place, it nevertheless suggests a continual volume of news throughout a day, irrespective of whether trading takes place. This fact will be useful when the return distributions are compared across different types of news.

3. Return Volatility and News

A basic tenet of financial economics is that asset prices change in response to unexpected fundamental information. Section 2.B describes a wide variety of news types from unidentified to identified. 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 seminal 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 that showed little relation between stock prices and news (used as a proxy for information), the finance literature has provided many analyses which demonstrate little 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. The basic 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 2,

[14] *Ravenpack* provides a contrast with identified news days being more common during the day than overnight. Specifically, identified news is captured by 146,269 firm-day news observations versus 88,247 firm-overnight news observations, representing 12.6% and 7.6% respectively. In general, these percentages are higher than those implied by *TSS*, suggesting *Ravenpack* is somewhat more successful at capturing identified news.

however, one of the issues with this literature may be the inability to recognize which news is relevant or not.

Three recent and related papers to ours are Griffin, Hirschey and Kelly (2011), Engle, Hansen and Lunde (2011) and Neuhierl, Scherbina and Schlusche (2013). Griffin, Hirschey and Kelly (2011) cross-check global news stories against earnings announcements to try and uncover relevant events. Engle, Hansen and Lunde (2011) utilize the Dow Jones Intelligent Indexing product to match news and event types for a small set of large firms. Neurhierl, Scherbina and Schlusche (2013) document significant stock price responses to a wide array of corporate press releases. 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.

In this section, we tie the literature documenting 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, starting with French and Roll's (1986) highly cited paper. The motivation of that paper was to better understand whether volatility is caused by public information, private information revealed through trading, or pricing errors by investors. (For a theoretical discussion, see Black (1986), Admati and Pfleiderer (1988), Foster and Viswanathan (1990, 1993), Madhavan, Richardson and Roomans (1997), and the survey by Madhavan (2000), among others.)

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 among 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., 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 large 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. For example, the null would suggest that of the 700 thousand no news days, 140 thousand would coincide with returns at the bottom and top 10%, 280 thousand would coincide with returns at the following 40%, and so forth. 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. In particular, they are 6.6% less likely to be extreme relative to the unconditional. This is consistent with the notion that news coverage proxies for information arrival. 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 1.6% more than the expected fraction of our defined "extreme" days. This is an important finding in the context of this paper. *Ex-ante*, one might have imagined that large price moves would have generated “news” stories, but this result shows that there is no mechanical relation between news and firm volatility.

Second, in sharp contrast to these results, we find that identified news days are 32.5% more likely to coincide with the extremes - the bottom 10% and top 10% of return days. Thus, while we might expect under independence to have 15,983 identified news stories in the extreme tails bucket, we actually observe 21,177 news stories in that bucket. That is, identified news days, but not unidentified news days, are much more likely to be extreme return days. Third, this pattern is much more pronounced for high intensity news days; these days are 78.3% more likely to coincide with extreme returns days.

To coincide with the existing literature, as a more formal look at the data, 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 high intensity 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 deviations on unidentified news days to no news days, and the

ratio of squared deviations on different types of identified news days to no news days.[15] For example, if both unidentified and identified news days have no additional effect on stock volatility, then we should find that these ratios are 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 is close to one (i.e., 1.20) while the variance ratio of identified news days exceeds two (i.e., 2.15). That is, the median stock exhibits return variance on identified news days that is 2.15 times the variance of no news days. 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 high intensity days, with 3.72 times the variance ratio. As additional evidence, Figure 1 depicts the distribution of these ratios across the 672 stocks for which these ratios are available (out of 791), winsorized at 10.[16] As evident, the ratios are not distributed around one for neither unidentified nor identified news days. However, the 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. The results are even stronger. The second-to-last column reports the ratio of median variance ratios of unidentified and identified news to no news days. While the ratio for unidentified news is 1.15, the variance ratio for identified news is 2.52, or 17.2% higher than *TSS*. Again, because the methodologies are quite different in nature, the results of Table 2 and Table A.2, panel A provide some comfort that (i) the methods capture relevant firm specific news, and (ii) this relevant news moves stock prices.

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

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

this news classification to revisit some well-known conclusions about the predominant role of private information arrival on stock return volatility.

In their seminal paper, French and Roll (1986) study 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. French and Roll 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 similarly during trading hours and overnight (i.e., 8.6%, 99,959 of 1,162,221, versus 6.5%, 75,162 of 1,162,221, respectively).

Second, appealing to behavioral finance, trading itself generates noise and higher volatility. Supply and demand shocks, possibly weakly related to fundamentals, affects prices through elastic supply and demand curves. Third, private information, not public information, is 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), Ito, Lyons and Melvin (1998), Barclay and Hendershott (2003), Madhavan, Richardson and Roomans (1997), 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 public information to be an important component of price variability. Consistent with these studies, in this subsection, we reexamine the results of Table 2, 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 high intensity 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 2.30% versus 1.33% overnight, that is, 76% 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.

This conclusion is further supported by two additional facts. First, return variances are relatively higher during trading days with no news, that is, on days in which there is no discernible public information. Specifically, median return volatility during trading hours versus overnight is 2.22% versus 1.14%, respectively; that is, 95% higher on no-news days compared to 76% on all days. Second, even on days with news, those typically associated with public information, return variances are 72% higher during trading hours (i.e., 2.39% versus 1.39%) if the news is unidentified.

Tables 2 Panels B and C, however, reveal a different story when the news can be identified, and especially so when the news is of high intensity. Specifically, on identified news days, the median trading day volatility is 2.89% versus overnight volatility of 2.05%, in other words, only 41% higher on identified versus 72% for unidentified news days. Equally important, the identified news median volatility of 2.05% overnight is close in magnitude to the volatility during trading hours on no-news days (i.e., 2.22%). 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 high intensity news. In particular, for high-intensity news, overnight volatility is similar to trading hours volatility, i.e., 2.72% versus 2.91%. 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 variance ratios of returns between various news types and no news, overnight and during trading hours. Specifically, overnight, the median

variance ratio of returns on unidentified news, identified news and high-intensity news days to no news days is 1.38, 2.71, and 5.55, respectively. This contrasts with significantly lower variance ratios during trading, i.e., 1.14, 1.59, and 2.11, 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. This is because variance ratios are lower during trading hours when private information can be revealed through trading in contrast to overnight. On the other hand, on identified and high intensity news days, the variances are 59% to 113% 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 there is, by definition, no trading. Overnight, the stock return variances are 171% to 455% higher on identified news and high intensity 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. On high relevance (identified, in our context) news days, the median trading day volatility is just 4.7% higher than overnight volatility (i.e., 2.70% versus 2.58%), and both are greater than the volatility during trading hours on no-news days (i.e., 2.04%). This translates to high variance ratios overnight for identified news, i.e. 6.31, compared to just 1.63 during trading hours.

In Section 4 we build on these results by suggesting a simple model that allows us to *quantitatively* identify the relative importance of public information on overall variance.

B. $\mathbf{R^2}$

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. This has been an important topic in both the theoretical and empirical finance literature (e.g., French and Roll (1986), Black (1986), Roll (1988) and Cutler, Poterba and Summers (1989)). A seminal paper on the question of whether stock prices reflect fundamental information is Roll (1988). In that paper, Roll

(1988) argues that once aggregate effects have been removed from a given stock, the finance paradigm would imply that the remaining variation of firm returns 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. Roll (1988) finds little discernible difference. Thus, his conclusion is that it is difficult to understand the level of stock return variation. 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 the other expanding Roll's (1988) puzzle even more. In particular, we find that when one can identify news, the news matters, but that there are not near enough identified news events (and aggregate movements) to explain stock returns.

The documented stylized fact in Table 2, that variances are higher on days in which we can identify important events and on days with high-intensity news, supports a relation between prices and fundamentals. As a more formal analysis, we reproduce the aforementioned Roll (1988) analysis for our setting. Table 3 reports results for a reinvestigation of the R^2 analysis of Roll (1988). 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 high intensity news.[17] All R^2 are adjusted for the number of degrees of freedom.

The results in the top part of Table 3 report median R^2 across. Consider the median calculations for the one-factor model. The R^2s are similar on no news and unidentified news days (i.e., 31.8% vs. 29.0%). The magnitude of the R^2s and similarity of these numbers between no news and news days (albeit unidentified) are consistent with Roll's puzzling

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

results. However, R^2s are much lower on identified news day, i.e., 15.7%. 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 2.0, in sharp contrast to Roll's results. Similar to the results from Table 2 with respect to variance ratios, the results are even more pronounced on high intensity news days, with R^2s lower, i.e., 10.7%.

Roll's original model-based null hypothesis, dramatically refuted empirically in his 1988 work, was that the performance of a market model, as measured by R^2, should be much worse during days on which firm-specific information arrives, compared with days when no such information arrives. In contrast to Roll's results, our results do lend support to this conjecture, since we are able to better proxy for firm-specific information arrival days using event identification.

Our results appear to be robust to the pricing model. For example, the results are analogous for the four-factor model that, along with the market, includes the book-to-market, size and momentum factors. In particular, the ratio of median R^2 between no-news and identified news days is only slightly lower (1.96 versus 2.02), and the R^2s between no-news and unidentified days is again similar.

That said, even though the drop in R^2s from no news days to identified news days is impressive, there is still a substantial unexplained variability in stock returns. Of course, while these low R^2s may be partially due to model or measurement error, one of the major puzzles of Roll (1988) remains unexplained. That is, on days in which there is no news on the Dow Jones wire, either identified or unidentified, the market (or four factor regression) still only explain 31.8% (38.6%) of the variation. This suggests a behavioral explanation or considerable stock return variation due to private information being impounded in prices via the trading process. To better understand the behavioral implications of the relation between identified news types and stock returns, below we try to partially 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 and, as we shall see, deepens the excess volatility puzzle.

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. In order to evaluate this explanation further, we run factor regressions using trading hours returns and overnight returns, conditional on various types. These results are reported in Table 3, columns 6-11.

On the one hand, the results strongly support the hypothesis that when important public information is identified, this information matters for stock prices. During closing hours, that is, when no trading takes place, R^2s for identified news and high intensity news days are 11.3% and 19.1%, respectively compared to, during trading hours, 17.1% and 14.0%. That is, when we isolate to a period with highly relevant public information without either private information trading or noise trading taking place, the explanatory power of aggregate factors drops in most cases.

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 cannot be a source for unexplained variability, conditioning on either no news or unidentified news, R^2s are only 25.2% and 22.7% respectively. More important, these R^2s are actually slightly lower than the R^2s of 26.4% and 23.8% during trading hours. This latter result fine-tunes and deepens the challenge for rational pricing. The results in this paper show that relevant, public information is important for explaining stock price variability. The problem is that once that information is accounted for, and by construction in close to open returns we move away from a trading or volume-based explanation, it is not clear what rational possibilities remain.

4. 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. 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. We compare the results from this decomposition both cross-sectionally and across years.

By definition, daily returns can be broken up into two components: trading hours returns and overnight returns. These returns can then be further separated into components conditional on identified news versus no news/unidentified news days. Equation (1) represents this breakdown:

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

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; and $\sigma^2_{TRDNG:NoNews,jt}$ is the trading day return variance of firm j at time t conditional on no relevant information. 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.

Since the goal is to quantify the contribution of public information to stock return volatility, it is not sufficient to simply estimate the return volatility on days with public information. This is because return volatility exists on days without any relevant information. To address this, we make the assumption that the return volatility due to public information is independent of other sources of return volatility. Under this assumption, equation (1) can be rewritten as a regression equation:

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

where $I_{OVRNT:News,jt}$ 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, as well as break up the sample by firm characteristics (i.e., volume, coverage, size, value, momentum and industry).

The overall variance contribution of news is a product of (1) the impact of news upon arrival, and (2) the intensity of news 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 4A and 4B respectively provide estimates of the coefficients from regression equation (2) and the economic interpretation of those coefficients. The coefficient $\beta_{OVRNT:News}$ can be interpreted as the incremental variance contribution coming from public information during closing hours, while $\left(\beta_{TRDNG:News} - \beta_{TRDNG:NoNews}\right)$ represents the variance contribution coming from public information during trading hours. The first column represents regressions of raw returns, the second column excess returns over the market return, and the third through sixth columns cover idiosyncratic returns from a market model regression with various combinations of fixed effects. The results are robust to all these specifications. We therefore focus on the idiosyncratic volatility estimates provided in the last column, which includes firm and date fixed effects.

First, and foremost, the return variance contribution of news is positive and large. For example, overnight, variance on identified news relative to no news days increases by 5.10, relative to unconditional mean of 1.90 (Panel B, last column). The incremental contribution of news is similar during trading hour, with news delta of 4.63, while the unconditional level of variance during trading hour is much higher (4.86). Thus, the relative increase of news is more pronounced overnight.

Second, as shown in Table 1 and repeated here in Table 4B, the fraction of overnight news days (8.60%) is marginally higher than those of trading hours (6.47%). Coupled with the incremental news contribution result discussed above, we show that the contribution of news to overnight volatility is much greater compared to its contribution during trading hours. Specifically, 23.1% of overnight return volatility is explained by news even though only 6.2% of the days have news.

Finally, it is worth comparing 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. While the qualitative results remain unchanged, the analysis based on *Ravenpack* suggests an even larger role for news on variance, especially overnight. First, the contribution of relevant news is relatively higher overnight than during trading hours using *Ravenpack* (6.34 versus 4.01%) rather than using *TSS* (5.10 versus 4.63). Second, *Ravenpack* identifies considerably more relevant news days than does *TSS*, especially overnight - specifically, for Ravenpack, 12.6% and 7.6% overnight and during trading hours, respectively, versus TSS' 8.6% and 6.5%. Points 1 and 2 combined lead to the startling result that 42.0% of overnight return volatility is explained by news compared to just 6.3% of trading day return volatility. The latter finding is quite consistent with *TSS*, but the former result suggests that *Ravenpack* is more adept at uncovering relevant news events released overnight. Because there are no prices observed during these hours, it is likely that the news identified by *Ravenpack* is tied to firm specific news.

Beyond the relative magnitude of news variance contribution during trading hours and overnight, the results using both TSS and Ravenpack suggest that a sizeable fraction of idiosyncratic variance can be traced back to the arrival of relevant firm-level public news. This suggests that public news have a more significant role than previously considered. It also provides an approach for decomposing return variance described by equations (1) and (2) can be employed quite generally to better understand the role of relevant public information. In the next few subsections, we explore this idea by looking at firm characteristics, industries and year effects.

A. Firm Characteristics

Overnight, private information driven trading cannot be the driver of variance. Thus, a natural interpretation of the fraction of overnight variance which is news-driven is one of a proxy for mispricing. Given the vast literature in finance that debates the sources of cross sectional differences in returns (e.g., Fama and French (1993), Jegadeesh and Titman (1993) and Carhart (1997)), we ask whether cross sectional characteristics are correlated with this proxy. In particular, we follow standard sorts and break the sample into large versus small firms, high book-to-market versus low book-to-market, and winners versus losers (i.e., momentum). Two additional characteristics that are of interest are volume and coverage. Since volume and coverage are highly correlated with size, we orthogonalize them by pre-sorting on size.

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., 6.99 versus 4.51 overnight and 7.53 versus 3.74 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., 10.54% versus 4.69% overnight and 7.80% versus 3.80% during trading hours). While these effects offset, the dominant factor is the frequency of news, with total return variance being explained by public news being equal to 26.49% and 6.71% for large firms versus 16.35% and 5.15% for small firms, during overnight and trading hours, respectively.

Firm size is highly correlated with both firm volume and with news coverage. To analyze these two characteristics separately, we adjust for firm size and focus on volume and news coverage separately. That is, we assign firms into volume and news coverage bins conditional on their size bins. We confirm that the double sort indeed results in volume and coverage being independent of size by computing the cross-sectional correlation of the orthogonalized characteristics with size. The empirical correlations are 0.013 and 0.017 for volume and coverage, respectively.

First consider volume. In contrast to the size characteristic above, high volume firms have both higher incremental news deltas overnight (e.g., 6.50 versus 3.00), as well as higher frequency of identified news (11.81% versus 5.45%). This leads to almost double the return variance explained by public news overnight (e.g., 28.45% versus 14.76%).

Size-adjusted news coverage results are striking, with 29.28% of the overnight return variance being driven by high news covered firms compared to 13.48% for less covered firms. Indeed, news intensity rises for high coverage firms, by construction -- high coverage firms have identified news on 13.69% of the days compared to 3.80% for low coverage. What is not by construction is that the news deltas that underlie these patterns are similar for low and high coverage firms, 5.20 versus 4.96. Note that if the extra coverage was spurious, it would have diluted news deltas.

Viewing these results in light of the fraction of total overnight variance attributable to news as a proxy for mispricing, the analysis offers some interesting observations. Small stocks, even in our universe of S&P500 firms, are substantially more prone to mispricing compared to large stocks – large stock overnight volatility attributable to news is 26.5%, 62% higher than the 16.4% for small stocks. This is consistent with the view that small stocks, all else equal, tend to be more mispriced than large ones (e.g, Berk (1995)). Likewise, we find substantial differences between low versus high volume stocks, with the ratio of news driven variance being 93%, and low versus high coverage ratio being 117% higher. To the extent that noise trading risk is priced (De Long, Shleifer, Summers and Waldmann (1990)), these results are broadly consistent with premia accrued to investors holding small, illiquid and low visibility stocks (see Fama and French (1993), Pastor and Stambaugh (2003) and Fang and Peress (2009) respectively). We find only moderate differences for winner versus loser momentum stocks and for value versus growth stocks. This non-result is in keeping with the noise trading premia explanation above.

The large differential in return variance contribution between large and small firms, high and low volume stocks, and high and low coverage suggests that a double sort would demonstrate even more dramatic results. Table 5B presents the same analysis for three different sorts – (a) large firms, high volume, (b) large firms, high news coverage, and (c) high volume, high coverage. The results are fairly robust across all three sorts. Compared to

the results in Table 5A, the double sort leads to return variance contributions ranging from 32.75% to 34.50% for overnight hours, and 7.83% to 8.98% for trading hours. While the dominant factor is the frequency of identified news being captured by these sorts, it is also the case that the incremental variance delta of the news also tends to be higher. 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.

As a final comment, note that similar to our breakdown of returns into overnight and trading day returns, Lou, Polk and Skouras (2015) provide a typical cross-sectional asset pricing analysis for these two distinct periods. They find that the momentum factors "works" during overnight hours while book-to-market and size "work" during trading hours. One potential explanation for momentum is underreaction to news. Given that we find identified news is more likely to be impactful during overnight hours, it is interesting that momentum only works during these hours. Nevertheless, Lou, Polk and Skouras (2015) find no real statistical difference across news versus no-news months, as defined by months including an earnings announcement or news coverage in Dow Jones Newswire. In this context, the findings in table 5A find little difference in the return variance contribution of identified news between growth and value firms (as represented by book-to-market) and winners and losers (as represented by momentum). This suggests that identified news is perhaps not the determining factor describing these characteristics or risk factors. Nevertheless, an interesting extension of Lou, Polk and Skouras (2015) would be to narrow such news coverage to relevant news identified by a textual methodology like *TSS* or *Ravenpack*.

B. Year

The sample period covers a number of unique, highly volatile episodes, including the burst of the so-called "dotcom" bubble during 2000 and 2001, and the emergence of the financial crisis, in particular, the bankruptcy of Lehman Brothers during the Fall of 2008. While these events are not firm specific, nevertheless, it is interesting to analyze their impact on the

variance decomposition of firm specific news. In Table 6 we follow the methodology outlined above, estimating the model for each year separately.

The percentage of identified news days during trading hours and overnight is fairly consistent over the decade-long period. For example, the average percentage varies in a tight range from 6.64% to 9.68% overnight versus 4.23% to 7.99% for trading hours. Moreover, for every year, there are a greater number of identified events overnight with a typical difference being approximately 1%-2%. Of some importance, there is nothing notable *per se* about the highly volatile "crisis"-like periods at the beginning and end of the sample.

In contrast, for these periods, there are significant differences in variance delta of news over time. The incremental contribution of relevant news is considerably higher during 2008 with an incremental variance contribution of 14.33 and 13.14 for overnight and trading hours, respectively. This compares to just 2.15 and 2.04 respectively for the period 2003-2007. This result suggests that it is neither the quantity of information, nor the timing of the information during the day, but the added firm-specific fundamental uncertainty during extreme periods, which drives returns. That said, other than identified news during the 2008 crisis, which explains 33.60% of overnight variance and 9.38% of trading-day variance, the other extreme periods do not explain a greater portion of total return variance. This is undoubtedly because aggregate volatility is higher during these periods, thus diluting the relative impact of fundamental news.

C. Industry

While the literature has gone the way of characteristics, perhaps a more natural way to distinguish between stocks and the way information is impounded in their prices is to group them by sectors. Intuitively, firm economics that are driven primarily by aggregate fundamentals such as commodities might be expected to be relatively less dependent on firm specific news. These industries include nondurables, energy and utilities, among others. Table 7 presents the findings. The results are mixed. On the one hand, firm specific news in both energy and nondurables provides little incremental variance contribution, 0.45

and 0.94, respectively, for overnight and 0.83 and 3.17 for trading day hours. These numbers lead to just 2.21% (0.83%) and 5.23% (4.90%) of variance for the energy and nondurables industry during overnight (trading day) hours being explained by relevant news. On the other hand, results for the utilities' industry are similar to that of the overall cross-section with 22.68% of overnight and 7.06% of trading day return variance coming from relevant news.

The two most responsive industries in terms of the relation between relevant news and stock return variance are financials and telecommunication, namely 28.31% and 30.80% explanatory power during overnight hours. The reasons for these two industries being an exception are, however, quite different. For telecommunication companies, likely due to the extraordinary changes occurring during the internet and mobile revolution, the defining characteristic are the fraction of news days, i.e., 16.98%. In contrast, the incremental contribution for financials is the highest of any industry, 8.29. As shown above in Section 4B, the year 2008 is an outlier. Given the financial crisis, and its impact on large financial institutions that make up the S&P500, it is not a far reach to link the results for financials in Table 7 with that of 2008 in Table 6.

A more systematic depiction of this intuition is obtained by examining, across industries, the relationship between average residual variance and the fraction of variance that is explained by news, plotted in Figure 2. A tight relationship is apparent with the R-squared of the cross sectional relation being 54%. That is, industries with higher residual variance (e.g., Business Equipment, Financial and Telecom) are also the ones with the highest fraction of explained news variance. The slope coefficient of the cross sectional relation is 0.11, implying that as the residual variance doubles, the level of news-driven variance rises by 11%. All in all, our ability to attribute variance to identified firm-level public news provides useful in explaining observed differences across sectors.

5. Conclusion

In this paper, we provide a methodology that allows us to isolate the portion of return variance due solely to the arrival of relevant firm-level public news. The bottom line is that, when relevant news can be identified, there is a much closer link between it and stock prices. 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 20%-40% overnight and 6% during trading hours; and model R^2s that are no longer the same on news versus no news days, but now are 16% versus 31%.

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, especially on the behavioral side. 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, there is a vast literature in behavioral finance arguing 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.

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

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, *The Journal of Finance* 41, 529–543.

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

Carhart, M. M., 1997, On persistence in mutual fund performance, *The Journal of Finance* 52(1), 57-82.

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

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

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, *The 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*.

Devitt, A., and K. Ahmad, 2007, Sentiment Polarity Identification In Financial News: A Cohesion-Based Approach, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 984–991.

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

Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33(1), 3-56.

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.

Fang L., and J. Peress, 2009, Media Coverage and the Cross-Section of Stock Returns, *The 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., B. Rosenfeld, R. Bar-Haim, and M. Fresko, 2011, The Stock Sonar Sentiment Analysis Of Stocks Based On A Hybrid Approach, *Proceedings of the Twenty-Third Innovative Applications of Artificial Intelligence Conference*, 1642–1647.

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.

Foster, F. and S. Viswanathan, 1993, The effect of public information and competition on trading volume and price volatility, *Review of Financial Studies*, 6 (1): 23-56.

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, The Arrival of Information and the Reaction of Traders, *Journal of Financial Economics* 17, 5–26.

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

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.

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

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, N., and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *The Journal of Finance* 48(1), 65-91.

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*.

Kothari, S. P., X. Li, and J. E. Short, 2009, The Effect Of Disclosures By Management, Analysts, And Financial Press On The Equity Cost Of Capital: A Study Using Content Analysis, *Accounting Review* 84, 1639–1670.

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

Lavrenko, V., M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, 2000, Mining Of Concurrent Text And Time Series, *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 37–44.

Li, F., 2010, The Information Content Of Forward-Looking Statements In Corporate Filings A Naive Bayesian Machine Learning Approach, *Journal of Accounting Research* 48, 1049–1102.

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

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.

Madhavan, A., 2000, Market microstructure: A survey, *Journal of Financial Markets*, 3/3, 205258.

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.

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.

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

Pastor, L., and R. F. Stambaugh, 2003, Liquidity Risk And Expected Stock Returns,*Journal of Political Economy* 111(3), 642-685.

Richardson, M. and T. Smith, 1994, A Direct Test of the Mixture of Distributions Hypothesis: Measuring the Daily Flow of Information, *Journal of Financial and Quantitative Analysis*, 29, 101-116.

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, 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.

Tables

Table 1: Summary Statistics

Panel A: DAY

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Total	1,245,709	791	3.6	325	58
No News	705,430	790	NA	NA	NA
Unid News	380,450	791	2.6	329	49
Iden News	159,829	790	6.1	316	81
High Inten News	37,151	740	11.7	320	78

Panel B: OVRNT

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Total	1,162,221	745	2.9	323	52
No News	765,278	745	NA	NA	NA
Unid News	296,984	745	2.2	329	44
Iden News	99,959	744	4.9	305	77
High Inten News	10,273	554	10.9	310	86

Panel C: TRDNG

	# Obs.	# Tickers	# Articles (daily)	# Words (per art.)	# Relv. Words (per art.)
Total	1,162,221	745	2.4	337	65
No News	870,147	745	NA	NA	NA
Unid News	216,912	743	1.9	339	54
Iden News	75,162	730	4.0	333	97
High Inten News	7,831	539	9.1	329	88

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 a corporate event, High Intensity News days are identified news days on which more than two different corporate 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%	2.64	791	1.16**
No News	-6.6%	0.5%	2.8%	2.38	781	
Unid News	1.6%	-0.4%	-0.4%	2.66	764	1.20**
Iden News	32.5%	-5.5%	-10.7%	3.51	681	2.15**
High Inten News	78.3%	-14.1%	-25.0%	4.53	401	3.72**

Panel B: TRDNG

	Return Rank			Stock SD and Variance		
	20% Extreme	40% Moderate	40% Low	Med SD	N Tickers	Var Ratio
Total	1.0%	-0.5%	0.1%	2.30	745	1.04**
No News	-3.0%	0.0%	1.5%	2.22	745	
Unid News	5.4%	-1.2%	-1.5%	2.39	658	1.14**
Iden News	34.4%	-5.5%	-11.7%	2.89	546	1.59**
High Inten News	97.0%	-17.8%	-30.8%	2.91	119	2.11**

Panel C: 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%	1.33	745	1.22**
No News	-4.6%	0.4%	2.0%	1.14	739	
Unid News	4.9%	-1.5%	-1.0%	1.39	704	1.38**
Iden News	32.2%	-6.6%	-9.5%	2.05	585	2.71**
High Inten News	97.7%	-20.7%	-28.2%	2.72	126	5.55**

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 3: R^2s – Firm-level Regressions

			DAY				OVRNT			TRDNG	
	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	791	27.0%	1.18**	32.7%	1.18**	745	20.3%	1.24**	745	25.1%	1.05**
No News	774	31.8%	1.00	38.6%	1.00**	734	25.2%	1.00	744	26.4%	1.00
Unid News	721	29.0%	1.10**	35.9%	1.08**	648	22.7%	1.11**	613	23.8%	1.11**
Iden News	597	15.7%	2.02**	19.6%	1.96**	501	11.3%	2.23**	440	17.0%	1.55**
High Inten News	262	10.7%	2.97**	14.4%	2.68**	47	19.1%	1.32**	30	14.0%	1.89**

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 4: News Variance Contribution

Panel A: Variance Regressions

Dependent Variable	Ret^2	Res^2	Eps^2	Eps^2	Eps^2	Eps^2
$I_{OVRNT,News}$	5.284	4.916	4.809	5.084	4.854	5.100
	[0.318]***	[0.303]***	[0.296]***	[0.125]***	[0.121]***	[0.124]***
$I_{TRDNG,News}$	8.879	7.588	7.407	7.689	7.512	7.732
	[0.316]***	[0.297]***	[0.291]***	[0.141]***	[0.139]***	[0.141]***
$I_{TRDNG,NoNews}$	4.243	3.225	3.099	3.105	3.096	3.104
	[0.044]***	[0.041]***	[0.040]***	[0.050]***	[0.050]***	[0.050]***
Constant	1.972	1.567	1.485	0.000	0.000	0.000
	[0.032]***	[0.031]***	[0.031]***	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
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	8.60%	8.60%	8.60%	8.60%	8.60%	8.60%
TRDNG, frac of News days	6.47%	6.47%	6.47%	6.47%	6.47%	6.47%
OVRNT News Δ	5.28	4.92	4.81	5.08	4.85	5.10
TRDNG News Δ	4.64	4.36	4.31	4.58	4.42	4.63
OVRNT News Var Contribution	18.73%	21.25%	21.78%	23.03%	21.99%	23.10%
TRDNG News Var Contribution	4.60%	5.56%	5.73%	6.10%	5.87%	6.15%

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.

Table 5: News Variance Contribution and Firm Characteristics

Panel A: Cross-Sectional News Variance Contribution

	Size		Volume (⊥Size)		Coverage (⊥Size)		BM		MOM	
	Coeff.	SdErr.	Coeff.	SdErr.	Coeff.	SdErr.	Coeff.	SdErr.	Coeff.	SdErr.
I_{High}	0.500	0.110	1.513	0.101	0.930	0.094	0.268	0.113	-0.244	0.087
$I_{OVRNT,News}$	6.993	0.276	3.008	0.212	4.958	0.248	4.443	0.189	6.168	0.188
$I_{High,OVRNT,News}$	-2.483	0.309	3.496	0.261	0.238	0.287	1.372	0.277	-2.065	0.279
$I_{TRDNG,News}$	11.123	0.305	5.145	0.242	7.520	0.290	6.907	0.212	8.090	0.212
$I_{High,TRDNG,News}$	-4.653	0.344	4.238	0.297	0.327	0.332	0.801	0.314	-1.121	0.317
$I_{TRDNG,NoNews}$	3.594	0.084	2.005	0.069	2.842	0.068	3.030	0.075	3.318	0.077
$I_{High,TRDNG,NoNews}$	-0.862	0.105	2.289	0.100	0.563	0.100	0.044	0.110	-0.207	0.115
	Small	Large	Low	High	Low	High	Growth	Value	Loser	Winner
OVRNT News Δ	6.993	4.509	3.008	6.504	4.958	5.195	4.443	5.815	6.168	4.103
TRDNG News Δ	7.529	3.738	3.140	5.089	4.677	4.441	3.877	4.635	4.772	3.858
OVRNT (uncon. mean)	2.01	1.79	1.111	2.701	1.398	2.429	1.661	2.122	2.137	1.769
TRDNG (uncon. mean)	5.56	4.34	3.080	6.678	4.180	5.586	4.567	5.001	5.234	4.775
OVRNT, % News days	4.69%	10.54%	5.45%	11.81%	3.80%	13.69%	8.50%	8.18%	8.75%	8.59%
TRDNG, % News days	3.80%	7.80%	4.10%	8.88%	2.74%	10.42%	6.55%	11.77%	6.68%	6.43%
OVRNT: News Var.	16.35%	26.49%	14.76%	28.45%	13.48%	29.28%	22.73%	22.42%	25.25%	19.92%
TRDNG: News Var.	5.15%	6.71%	4.18%	6.77%	3.07%	8.28%	5.56%	5.63%	6.09%	5.19%

Panel B: News Variance Contribution of Large, High Volume and High Coverage Stocks

Characteristic	Large, High Vol	Large, High Cov	High Vol, High Cov
$I_{OVRNT,News}$	5.818	4.662	6.073
	[0.227]***	[0.200]***	[0.225]***
$I_{TRDNG,News}$	7.687	6.509	9.065
	[0.257]***	[0.227]***	[0.253]***
$I_{TRDNG,NoNews}$	3.584	3.023	4.045
	[0.118]***	[0.112]***	[0.121]***
Observations	738,584	729,172	778,234
R^2	0.002	0.002	0.003
Fixed Effects	Firm, Year	Firm, Year	Firm, Year
OVRNT (unconditional mean)	2.58	2.29	2.90
TRDNG (unconditional mean)	5.74	4.96	6.59
OVRNT, frac of News days	14.88%	14.88%	14.88%
TRDNG, frac of News days	10.96%	10.96%	10.96%
OVRNT News Δ	5.82	4.66	6.07
TRDNG News Δ	4.10	3.49	5.02
OVRNT News Var Contribution	33.62%	34.50%	32.75%
TRDNG News Var Contribution	7.83%	8.82%	8.98%

Panel A 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), 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. The Volume and Coverage dummies were determined based on median volume and coverage levels by year while orthogonolizing for size. Panel B repeats the analysis in Panel A for a subset of firms. Those that are either Large firms with High Volume (orthogonal to size), Large firms with High Coverage (orthogonal to size), or High Volume firms with High Coverage (orthogonal to size).

Table 6: News Variance Contribution by Year

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
$I_{OVRNT,News}$	5.575	5.664	8.383	2.523	1.984	2.102	1.990	2.169	14.331	4.486
	[1.413]***	[0.649]***	[1.432]***	[0.231]***	[0.196]***	[0.326]***	[0.317]***	[0.336]***	[1.996]***	[0.544]***
$I_{TRDNG,News}$	9.447	10.557	11.887	4.393	3.011	3.284	2.660	2.803	19.746	9.994
	[0.557]***	[1.151]***	[1.161]***	[0.217]***	[0.152]***	[0.422]***	[0.222]***	[0.366]***	[1.942]***	[0.962]***
$I_{TRDNG,NoNews}$	5.617	4.216	4.661	1.748	1.122	0.868	0.991	1.216	6.610	3.881
	[0.156]***	[0.130]***	[0.131]***	[0.045]***	[0.027]***	[0.127]***	[0.071]***	[0.058]***	[0.234]***	[0.131]***
Constant	2.655	2.045	2.047	0.913	0.616	0.709	0.608	0.728	2.731	1.797
	[0.135]***	[0.106]***	[0.074]***	[0.028]***	[0.021]***	[0.123]***	[0.039]***	[0.045]***	[0.175]***	[0.100]***
Observations	229440	230266	233000	236890	237356	237294	228086	227758	231670	231738
R^2	0.005	0.005	0.005	0.009	0.009	0.001	0.002	0.002	0.004	0.005
OVRNT (mean)	3.056	2.539	2.604	1.090	0.802	0.912	0.798	0.938	4.113	2.183
TRDNG (mean)	8.434	6.638	7.058	2.807	1.866	1.756	1.730	2.071	10.308	6.094
OVRNT %News days	7.20%	8.72%	6.64%	6.99%	9.39%	9.64%	9.56%	9.68%	9.64%	8.60%
TRDNG %News days	4.23%	5.94%	4.85%	5.52%	6.80%	7.38%	7.85%	7.99%	7.36%	6.79%
OVRNT News Δ	5.58	5.66	8.38	2.52	1.98	2.10	1.99	2.17	14.33	4.49
TRDNG News Δ	3.83	6.34	7.23	2.65	1.89	2.42	1.67	1.59	13.14	6.11
OVRNT: News Var	13.13%	19.44%	21.39%	16.19%	23.23%	22.22%	23.84%	22.39%	33.60%	17.67%
TRDNG: News Var	1.92%	5.68%	4.97%	5.20%	6.89%	10.16%	7.58%	6.12%	9.38%	6.81%

The top part of 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 return window with identified news and those without. Results are reported for each year in the sample separately. All regressions include firm and date fixed effects. The bottom part the table 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.

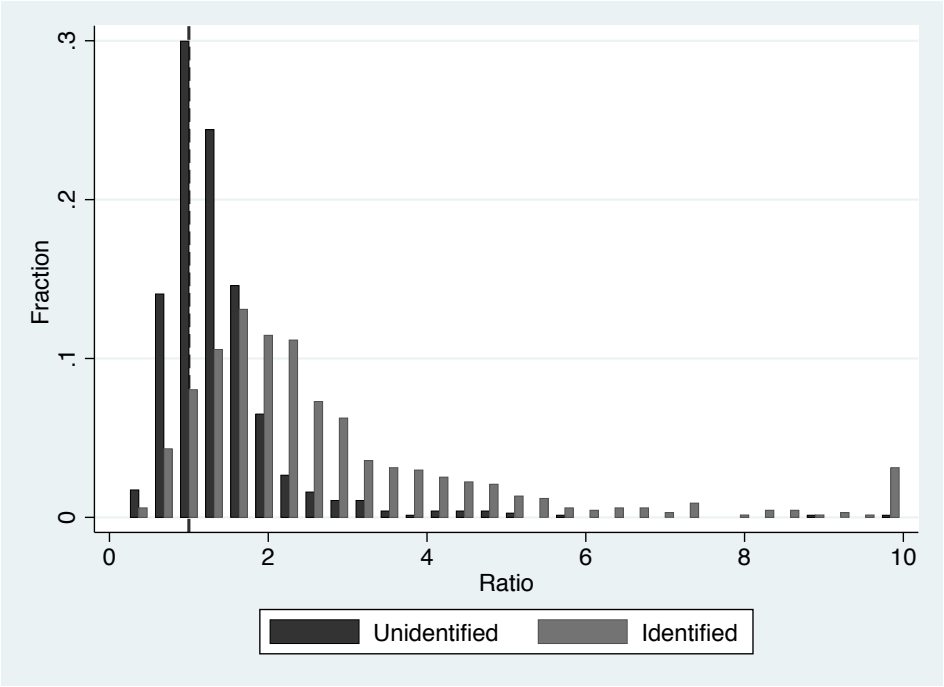
Table 7: News Variance Contribution by Industry

Industry	BusEq	Chems	Durbl	Enrgy	Hlth	Manuf	Money	NoDur	Other	Shops	Telcm	Utils
$I_{OVRNT,News}$	5.883	1.779	4.067	0.454	2.324	1.698	8.288	0.937	4.293	3.397	4.573	4.932
	[0.505]***	[0.653]***	[1.410]***	[0.323]	[0.373]***	[0.269]***	[0.852]***	[0.338]***	[1.555]***	[0.505]***	[1.663]***	[0.988]***
$I_{TRDNG,News}$	5.352	5.968	8.201	3.697	2.990	5.245	11.044	4.769	6.788	8.405	5.439	11.233
	[0.254]***	[1.368]***	[1.869]***	[0.438]***	[0.159]***	[0.282]***	[0.876]***	[0.439]***	[0.635]***	[1.122]***	[0.546]***	[1.795]***
$I_{TRDNG,NoNews}$	4.647	1.557	2.790	2.867	1.639	2.621	3.663	1.599	3.717	2.735	4.163	2.924
	[0.123]***	[0.118]***	[0.245]***	[0.284]***	[0.058]***	[0.079]***	[0.091]***	[0.192]***	[0.136]***	[0.097]***	[0.201]***	[0.125]***
Constant	2.029	0.874	1.613	1.344	0.983	1.141	1.863	1.109	1.533	1.392	1.745	1.033
	[0.102]***	[0.111]***	[0.159]***	[0.272]***	[0.049]***	[0.063]***	[0.066]***	[0.179]***	[0.098]***	[0.060]***	[0.101]***	[0.050]***
Observations	267,928	85,232	41,580	102,132	147,874	286,222	629,340	165,050	116,540	239,612	66,716	175,270
R^2	0.006	0.003	0.004	0.001	0.005	0.005	0.003	0.001	0.003	0.004	0.002	0.004
OVRNT (uncond. mean)	2.618	0.995	2.056	1.375	1.243	1.270	2.598	1.170	1.991	1.626	2.521	1.336
TRDNG (uncond. mean)	6.740	2.666	4.836	4.246	2.735	3.902	6.016	2.847	5.444	4.453	6.091	4.257
OVRNT, % News days	10.01%	6.81%	10.90%	6.70%	11.19%	7.59%	8.88%	6.53%	10.68%	6.87%	16.98%	6.15%
TRDNG, % News days	9.06%	5.33%	8.00%	4.23%	8.44%	5.33%	6.64%	4.40%	6.31%	5.75%	14.37%	3.62%
OVRNT News Δ	5.88	1.78	4.07	0.45	2.32	1.70	8.29	0.94	4.29	3.40	4.57	4.93
TRDNG News Δ	0.71	4.41	5.41	0.83	1.35	2.62	7.38	3.17	3.07	5.67	1.28	8.31
OVRNT: News Var	22.49%	12.18%	21.56%	2.21%	20.93%	10.15%	28.31%	5.23%	23.03%	14.36%	30.80%	22.68%
TRDNG: News Var	0.95%	8.82%	8.96%	0.83%	4.17%	3.59%	8.15%	4.90%	3.56%	7.32%	3.01%	7.06%

The top part of 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 return window with identified news and those without. Results are reported for each of the 12 main sectors separately. All regressions include firm and date fixed effects. The bottom part the table 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.

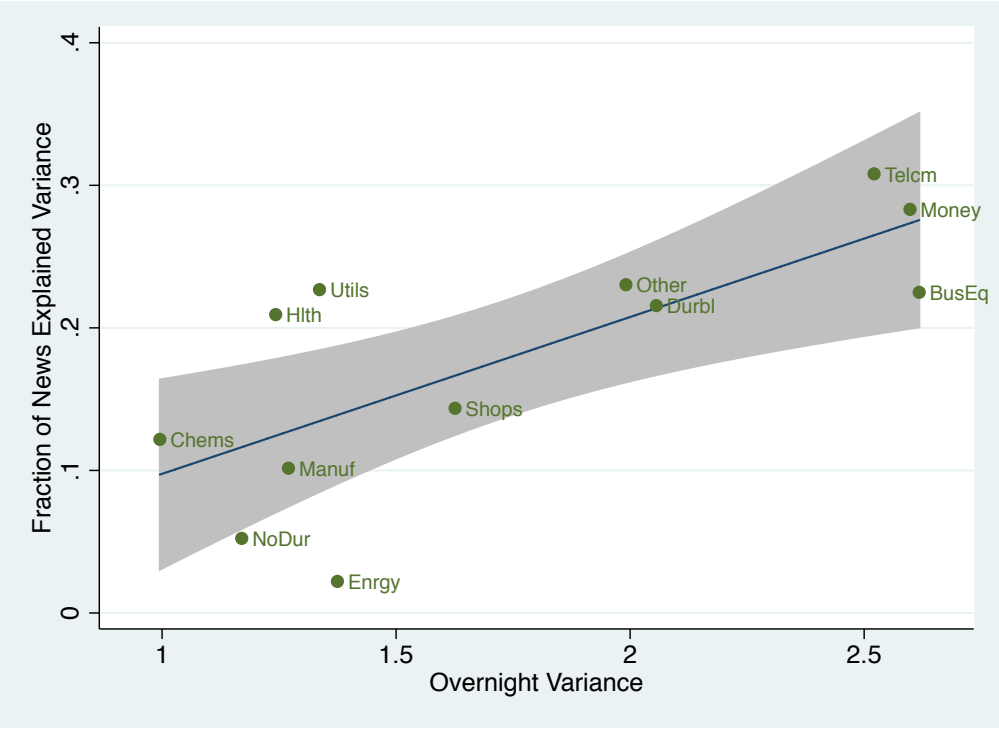
Figures

Figure 1: Distribution of Variance Ratios



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

Figure 2: News Variance Contribution Across Industries



The figure plots the average overnight firm residual returns squared, across industries, and the corresponding average fraction of news explained variance. The regression line (with 95% confidence interval) is added.

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.