News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies*

Yoontae Jeon† Thomas H. McCurdy‡ Xiaofei Zhao§

First draft: December 9, 2018

This draft: December 11, 2019

Abstract

A potential important source of jumps in stock returns can be material news events. In this paper, we collect 21 million news articles associated with more than 9000 publicly-traded companies and use textual analysis to derive measures summarizing those news. We find that measures of news flow content are significantly related to nonparametric measures of jumps. Moreover, by modelling the observable news process explicitly and jointly with a latent jump process, we find that news are important drivers of the jumps in stock returns. Consequently, we are able to enrich the economic content of the widely-used econometric models of jumps.

*We thank Martin Boyer, Stephen Figlewski (discussant), Luis Goncalves-Pinto (discussant), Ryan Israelson (discussant), Kristine Hankins, Tim Loughran, Sumudu Watugala, Dacheng Xiu and participants at the 2019 Risk Management and Financial Innovation Conference, 2019 European Financial Association Annual Meeting, 2019 Northern Finance Association Annual Meeting for helpful comments and suggestions. We also thank Tim Loughran and Bill McDonald for making their sentiment word list and master dictionary available.

†Ted Rogers School of Management, Ryerson University, yoontae.jeon@ryerson.ca

‡Rotman School of Management, University of Toronto, tmccurdy@rotman.utoronto.ca

§McDonough School of Business, Georgetown University, xiaofei.zhao@georgetown.edu

1 Introduction

There is a long history of asset pricing theory that links the quantity and quality of information flows to changes in asset prices. For example, information that results in a resolution of uncertainty about a firm's future prospects can result in a revision in the current price. According to this view, an important process affecting price movements is the news arrival process. There is a large literature concerned with the effect of news on stock returns.[1] Recent research links news flows to the distributional properties of stock returns. For example, in a study of 23 firms, Lee (2012) finds that jumps in returns are more likely to occur during scheduled news announcement times. Engle, Hansen, and Lunde (2012) show that public-news counts are related to the volatility of stock returns for 28 large firms. A source of price changes, in addition to public news, is the private information that is revealed through trading. Boudoukh, Feldman, Kogan, and Richardson (2018) use textual analysis to identify the effects of news that are associated with fundamentals as opposed to information revealed by trading; and test these alternative sources of return volatility.

Stock prices also exhibit large, discrete movements, typically labelled as "jumps". Jumps have been recognized as important for many financial and economic decisions, such as, portfolio re-balancing, derivative pricing, risk measurement and management. The intuitive idea that large movements in stock prices might be related to important information flows (such as earnings surprises) in the market has inspired many studies related to modelling jumps in stock returns, mostly treating information flow as latent. The basic jump-diffusion

[1] For example, research that has focused on the effect of newspaper articles on returns or volatility includes, among many others, Cutler, Poterba, and Summers (1989), Berry and Howe (1994), Mitchell and Mulherin (1994), Chan (2003), Antweiler and Frank (2004), Veldkamp (2006), Tetlock (2007), Tetlock, Saar-Tsechansky, and Macskassy (2008), Fang and Peress (2009), Tetlock (2010), Tetlock (2011), Kyle, Obizhaeva, Sinha, and Tuzun (2011), Manela and Moreira (2017), Bybee, Kelly, Manela, and Xiu (2019), and Ke, Kelly, and Xiu (2019).

model of Press (1967) and its many extensions can be applied to the effect of news flows on price changes.

This paper explicitly models the news processes associated with individual firms, rather than treating them as latent. The first part of our paper is a nonparametric analysis of firm-specific news. In particular, after collecting more than 21 million news articles associated with over 9000 publicly-traded companies from the Factiva database, we use *textual analysis* to derive measures summarizing those news, including news frequency, tone and uncertainty.[2] We then test to what extent those news variables explain nonparametric measures of jumps. Finally, we use those nonparametric measures of the news as inputs for our time-series modelling of firm-level news processes and their potential effect on stock return jumps. For example, we parameterize the news process as a mean-reverting process to capture the clustering of observable news arrival.

We study the top 20 firms with the most news coverage as well as $9,020$ firms sorted into large, medium and small firm size groups. The number of news articles across firms is heavily skewed towards large firms. For example, the group of small and medium firms are covered by 2.21 and 0.81 million news articles respectively, whereas the group of large firms are covered by 18.49 million news articles. The top 20 firms alone are covered by 3.16 million news articles. For example, the number of news articles covering IBM is $267,318$ between 1980 and 2012. Based on our textual analysis, the news tone is in general (slightly) negative, and more negative for large firms. Under the most stringent criteria of identifying daily jumps nonparametrically[3], there is one jump every 57 days; and the frequency of

[2] Our analysis builds on the word list provided by Loughran and McDonald (2011). See Loughran and McDonald (2016) for a recent survey on textual analysis in accounting and finance. There is a growing literature on using textual analysis to understand the content of news or regulatory filings. Also see, e.g., Gentzkow, Kelly, and Taddy (2018) and Kelly, Manela, and Moreira (2019).

[3] As discussed below, this is equivalent to identifying a daily jump if the absolute value of daily return exceeds 5.1 times the daily spot volatility.

jumps varies inversely across the firm size groups. The jump intensity gap across size groups becomes smaller once we relax the criteria for identifying the jumps.

The jump intensity is positively and significantly related to news arrival. The absolute value of the news tone is also positively and significantly related to jump intensity. The measures of news flow explain more variation in the jump intensity for large firms. We also conduct the same analysis for the top 20 firms and find that the R^2 of news flow explaining jump intensity can be as high as 15% (for Amazon). Conditional on at least one jump in daily returns, we link the jump size mean and volatility to news flow measures. The jump size mean is negatively related to news intensity, dominated by negative jump returns. Once we split the sample into positive and negative jump returns, we find that news intensity is positively related to positive jump returns and negatively related to negative jump returns. If we focus on the top 20 firms, we find that the R^2 of news flow measures explaining negative jumps is close to 26% and the R^2 of news flow measures explaining positive jumps is around 7%. For the jump size volatility, we find that both news intensity and the absolute value of the news tone are positively and significantly related to jump size volatility. The R^2 of news flow measures explaining jump size volatility is 12% for all firms and 24% for the top 20 firms. In general, our results suggest the measures of news flow explain an important fraction of variation in jump size distributions and news intensity plays an especially critical role in jump arrival.

To further explore the sensitivity of jump probabilities to news, we use cross-sectional analyses of the dependence of that sensitivity with respect to firm characteristic variables. Higher analyst coverage and a higher institutional ownership fraction both increase the senstivity of jump probabilities to news arrival, highlighting the importance of those channels with respect to quick incorporation of news into returns.

We also conduct an additional robustness analysis using the RavenPack news dataset and confirm that our findings are robust to using that alternative news database, as well as to using novel news instead of the entire set of news. We are also able to report some results for disaggregated news types using the RavenPack data. The Factiva data collected for this paper allows us to evaluate longer time periods.

Motivated by material news as a potential source of jumps in stock returns, Maheu and McCurdy (2004) proposed a GARCH-Jump model with time-varying jump arrival to capture the impact of unusual versus usual news events on stock returns for individual firms, as well as for various stock return indexes.[4] Due to the unavailability of comprehensive firm-level news data, Maheu and McCurdy (2004) did not model the news process directly. Rather they provided some examples for which days identified by their model as having high probability of at least one jump coincided with unusual and material news events. For example, in October 2000, IBM's negative earnings surprise of −18% led to a price change of −16.9%. This type of news surprise concerning expected future cash flows resulted in price changes well above normal and were better captured in their model by jumps rather than Brownian or normal innovations. The latter, less extreme movements in price, can be due to typical news events as well as liquidity trading and strategic trading as information disseminates.

The modelling of stock return jumps typically consists of two components: the frequency of jumps and the size of each jump. We expect the explicit news process to be an important

[4]There is a large literature on alternative component mixtures. For example, Chernov, Gallant, Ghysels, and Tauchen (2003) suggest that either a 2-component parameterization of stochastic volatility (SV) or a SV-jump-diffusion can capture the volatility dynamics. Other early examples of SV-jump-diffusion specifications with time-varying jump intensities include Bates (2000), Andersen, Benzoni, and Lund (2002), Pan (2002), Chernov et al. (2003) and Eraker, Johannes, and Polson (2003). Examples in a discrete-time setting with time-varying jump arrival include Bekaert and Gray (1998), Bates and Craine (1999), Neely (1999) and Das (2002), who allow a volatility factor or financial/macroeconomic variables to affect the jump intensity. Johannes, Kumar, and Polson (1999) consider a state dependent jump model which allows past jumps and observables to affect the jump probability.

driver of the latent jumps in stock returns. For example, one might expect the news frequency to be an important factor that drives the latent jump intensity of stock returns. In order to test the idea, we incorporate the firm-level news process into the jump arrival process component. By modelling the observable news process explicitly and embedding it into a latent jump process, we find that news are important drivers of the jumps in stock returns.

Specifically, we model the time series of the news processes jointly with latent jumps as an application of our findings. Building on the existing GARCH-class of model with compound Poisson jumps, we propose a new model where the jump intensity is driven by the level of the observed news count variable. The empirical fit of the model is overall comparable and mostly improves upon the benchmark GARJI model of Maheu and McCurdy (2004), indicating the power of our news measure in explaining the jump intensity dynamics. Out-of-sample analyses on predicting daily realized jump variation measures further support the importance of incorporating observable news measures into the parametric modelling. Overall, our parametric modelling analysis suggests potential benefit of explicit news measures when applied to risk management and portfolio allocation.

Our study differs from closely related papers in significant ways. Engle, Hansen, and Lunde (2012) focus on the impact of news count on stock return volatilities; whereas we analyze the contribution of news to stock return jumps. Using high-frequency returns for 23 stocks and a nonparametric method to identify jumps, Lee (2012) finds that there is a higher chance of having jumps during the scheduled announcement times. The firm-specific news considered by Lee (2012) include earnings announcements, analyst recommendation releases, and dividend dates. That research includes the effect of *scheduled* news on return jumps but does not model the news process itself. On average, the firms in her sample have 49 such news events every year over the sample period of 1993 to 2008. In contrast, our firms presented in Table 1 which are comparable to the firms in Lee (2012), have an average of $11,666$ news articles every year during our sample period from 1980 to 2012.

Both Lee (2012) and Engle, Hansen, and Lunde (2012) use a sample of large firms (23 and 28 firms respectively), whereas our sample is more comprehensive; we have over 21 million news articles associated with more than 9000 firms. Therefore, our paper sheds light on the impact of modelling news for large, medium and small companies. Although both Lee (2012) and Engle, Hansen, and Lunde (2012) focus on news arrival/count, we include not only the frequency of the news articles, but also the textual content measures of the news articles. Specifically, we investigate the impact of news on jump intensity, mean jump size, and volatility of jump size. Our analysis will have implications for a large class of parametric models on stock return jumps.

By modelling news processes explicitly, we are able to enrich the economic content of the widely used econometric models of jumps. Explicitly incorporating news processes in models of stock return jumps can potentially help the identification of jumps due to information arrival; and separate them from jumps due to other reasons such as liquidity and strategic trading based on private information. This may have broad implications for applications such as option pricing and risk management where stock return jump models are frequently used. For example, there is a literature on pricing jump risk for returns, such as Maheu, McCurdy, and Zhao (2013), and some other papers in the literature also have a focus on option pricing, for example, Pan (2002) and Christoffersen, Jacobs, and Ornthanalai (2012), have fit options and the underlying jointly with the focus on solving option pricing puzzles. These papers adopt parametric models for the underlying assets. Our analysis suggests that the class of option pricing models mentioned above can be enriched by incorporating the observed news measure explicitly.

Section 2 discusses our data collection and summary statistics of those data including news frequency measures, tone and uncertainty measures of the news based on textual analyses, and nonparametric measures of realized jumps in stock returns. Section 3 provides a detailed analysis of relationships between the realized jumps and the various measures of news flow, both with respect to jump frequency and statistics of the realized jump-size distribution. Section 4 introduces our joint parameterization of the dynamics of stock returns and news flow. The focus of this analysis is to embed the observed news flow in a time-series model of stock returns which features time-varying jump arrival. The estimation and out-of-sample performance of that parametric model are summarized in Section 5. The recent availability of the RavenPack news dataset enables us to report on some robustness analyses (with respect to novel versus all news) and some further results with respect to disaggregated news categories in Section 6. Section 7 provides some concluding comments.

2 Data

The stock return data are retrieved from CRSP. We collect all the news articles using Factiva. The news collection takes the following steps. First, we match CRSP/Compustat IDs with Factiva IDs, which allow us to obtain news articles for individual companies. We start with the public companies from the CRSP/Compustat merged database. For each of these companies, we search its CRSP/Compustat company name in Factiva to obtain the corresponding Factiva company name and company code. Second, for each of the companies with Factiva company code, we search for the total number of news articles available in the Factiva database. Our initial search started in July 2012, so we fix this date as the ending date of all our searches. Our news sample period is from January 1980 to July 2012. This round of search returns 9,020 companies with at least 1 news article in Factiva.

Third, we retrieve the news items for each Factiva firm. Due to the quota limitation of Factiva, we break down the searches for firms with large number of news items into searches by firm-month. We then use a set of Python scripts to organize all the news items retrieved from Factiva. For each news item, we are able to obtain the headline (title) and the first paragraph, the date of the news article, the media outlet where the news article is published, and the total number of words in the article.[5] We impose a Factiva filter to exclude the news articles discussing the market and stock price movements. This filter helps alleviating the concerns about reverse causality of stock return jumps triggering the news reports.

Table 1 presents the summary statistics of news count for our sample; including for all firms as a group, for firms in three size groups, and for the 20 individual firms selected from different industries with most news coverage which we label "*Top 20 News Firms*".[6] The total number of news articles for these firms altogether is 21.51 million. The number of news articles across firms is heavily skewed towards large firms.[7] The top 20 firms alone are covered by 3.16 million news articles. Not surprisingly, news articles concentrate in the post Internet period (post 2000). For all firms, close to 80% of the news articles are in the post 2000 period, although this coverage is smaller for small firms (62%) and much higher for the top 20 firms (with a few exceptions).

Table 1 about here

We analyze the textual content of the first paragraph of the news articles to obtain our

[5] We are also able to obtain the exact time stamps (up to seconds) for news articles from certain media outlets such as news wires.

[6] Note that these 20 firms do not necessarily correspond to 20 firms with the most number of news articles during the sample period as we selected firms with the most news coverage from various industries to avoid concentration of firms in a particular industry such as technologies.

[7] Size groups are defined by market capitalization at the end of the sample period. There are equal number of firms in three different size groups.

key textual measures (such as news tone and percent of uncertain words) using the word lists developed by Loughran and McDonald (2011), and Loughran and McDonald (2013).[8] Specifically, for the first paragraph of each news article, we calculate the news tone as the difference between the percentage of positive words and the percentage of negative words. We also calculate the percentage of uncertain words in each news article using the LM uncertain word list. Since most of our analyses are at daily level, we consolidate the article-level measures to generate corresponding daily measures. We use the number of words in each article as the weight to calculate the word-weighted daily news tone and daily percentage of uncertain words. The summary statistics of these daily textual measures are presented in Table 2. For the whole sample, the average daily news tone is -0.0315% and the average percentage of uncertain words is 0.2185%. Across the 3-size groups, large firms show more negative news tone and a higher percentage of uncertain words. This becomes more clear when examining the top 20 firms, whose news tone ranges from around -1% to -0.1% and the percentage of uncertain words ranges from around 0.2% to around 0.8%.

Table 2 about here

The realized jumps in stock returns are identified using the non-parametric approach from the literature (for example, Lee and Mykland (2008)).[9] Table 3 presents the summary statistics on the daily realized jumps by different firm groups. For example, $J99_{i,t}$ takes the value of 1 if there is a jump identified for stock i on day t, and 0 otherwise. The number next to J refers to the jump identification criteria and in this case (99) it corresponds to the

[8]Loughran and McDonald (2011) propose the word lists that are appropriate for business communication context. Following their paper, many of the recent papers analysing the textual content of news articles have adopted LM word lists when measuring sentiment or tone in the news articles. We follow this approach to avoid the subjectivity of creating our own word lists. See Loughran and McDonald (2016) for a survey on related studies.

[9]Alternatively, Bollerslev and Todorov (2011) estimate the market jump tail under the physical measure using high frequency intra-day data and estimate the risk-neutral counterpart from index options.

$99th$ percentile of the maximum return distribution from Lee and Mykland (2008). The total number of days with $J99 = 1$ for all firms is 351,374 (out of around 20 million days with non-missing returns), indicating that there is on average 1 jump every 57(=20/0.35) days. For the other measures ($J95$, $J_0 99$, $J_0 95$), we see more jumps on average as the criteria used to identify jumps become less stringent.[10] The frequency of daily jumps varies significantly inversely across the size groups. There is on average 1 jump every 80, 53, and 39 days for the large, medium, and small-size firm groups when $J99$ is used. The gap in jump frequency becomes smaller if we use less stringent criteria. For example, in the case of $J_0 95$, the jump frequency becomes 1 jump every 11, 9.5, 9 days for the large, medium, and small size firm groups.

Table 3 about here

3 Realized Jumps and News Flows

3.1 Realized Jump Intensity and News

We link the realized jumps to news flows. We start by using logistic regressions to examine how the probability of jump is related to the news flows measured by the news count, the (absolute value of) news tone, and the percent of uncertain words.

$$\text{logit}(p_{it}) = a + b_1 \times \text{NewsCount}_{it} + b_2 \times |\text{NewsTone}_{it}| + b_3 \times \text{UncWords}_{it} + b_4 \times |\text{ret}_{it-1}| + \epsilon_{it}, \quad (1)$$

where the dependent variable is the daily jump indicator variable. We also include the (absolute value of) lagged stock returns in the regressions. We expect the coefficient b_1 for

[10] Effectively, each of the four statistics $\{J99, J95, J_0 99, J_0 95\}$ identifies a daily return as a jump if the absolute value of the daily return is above $\{5.1024, 4.4881, 3.2283, 2.4565\}$ times the daily spot volatility.

NewsCount to be positive and statistically significant. In addition, we expect the absolute value of the news tone to be positively related to jumps.[11] Although a very high percentage of uncertain words can be related to negative jumps, a low percentage of uncertain words might not be related to positive jumps. Therefore, the relation between the percentage of uncertain words and the jump intensity would not be as strong as the news count and absolute value of news tone.

Note that when constructing the measure of news count we do not drop the news articles that are related to other news articles on the same day. By keeping all the news articles in the news count measure we aim to capture the importance of the news flows underlying the news articles. Undoubtedly this introduces some noise in the news count measure. Therefore, in the parametric modelling in Section 4, we allow measurement errors in the observed news count and use Kalman Filter to reduce the noise. The impact of related news articles on news tone and uncertainty is smaller as these are percentage measures.

Table 4 reports the results. Panel A reports the coefficient estimates for the whole sample and Panel B reports the standardized odds ratio associated with the corresponding variable. The news count is statistically significantly related to the probability of jump. The standardized odds ratio associated with the news count is 1.22, suggesting that one standard deviation increase in news count increases the odds of a jump by 22 percent. The absolute value of news tone is also positively and significantly related to the realized jumps with a standardized odds ratio of 1.02 (for $J99$). The effect of the percent of uncertain words is also positively related to realized jumps but the economic significance is weaker, with a standardized odds ratio of 1.01. In addition, the R^2 of the regression in column (1) is 1.4%. The patterns are in general similar when we used other measures of realized jumps

[11] This is because very positive news flow is likely to be related to positive jumps whereas very negative news is likely to be related to negative jumps.

in columns (2)–(4). In the appendix Tables A1-A3, we also repeat the analysis for large, medi um, and small size firm groups. For all three groups, news count is the most significant variable that relates to realized jumps. However, there are a couple of differences. In the large firms, the absolute value of news tone is more important than percent of uncertain words in explaining jumps whereas in medium and small firms, these two variables are of equal importance. Moreover, the news flows is much more important in explaining realized jumps for large firms. This is evident from the R^2 of 2.18% for large firms, followed by 1.36% and 0.79% for medium and small firms respectively (when $J99$ is used).

Table 4 about here

To better understand the cross-sectional variations in the impact of news flows on realized jumps, we also run the logistic regressions for each of the top 20 firms; the estimation results are reported in Table 5. For brevity, we only tabulate the results for $J95$. In general, the positive relation between news count and probability of jump holds for most of the firms (i.e., for 17 of the 20 firms). The magnitude of the coefficient estimates and the R-squared both have noticeable cross-sectional variations. For example, the R-squared is highest for Amazon (15%) and lowest for Bank of America (0.28%). The coefficient for news count is also highest for Amazon (0.048) and lowest for Cisco (0.001).

Table 5 about here

3.2 Realized Jump Size and News

Next we analyze how the jump size distribution is affected by news flows. We focus on the jump size mean and jump size volatility and expect the news content (i.e., news tone) to affect both moments significantly.

We focus on the observations of realized jumps to understand the impact of news on the first two moments of jump size distributions. For the jump size mean, we run the following regressions

$$r_{it}|\text{Jump} = b_0 + b_1 \times \text{NewsCount}_{it} + b_2 \times \text{NewsTone}_{it} + b_3 \times \text{UncWords}_{it} + b_4 \times \text{ret}_{it-1} + \epsilon_{it}, \quad (2)$$

where $r_{it}|\text{Jump}$ measures the realized jump returns. The above regression also implies the following relation between the jump size mean and news content measures

$$\begin{aligned} \mathbb{E}[r_{it}|\text{Jump=1}] \quad &= \quad b_0 + b_1 \times \mathbb{E}[\text{NewsCount}_{it}] + b_2 \times \mathbb{E}[\text{NewsTone}_{it}] \\ &\quad + b_3 \times \mathbb{E}[\text{UncWords}_{it}] + b_2 \times \mathbb{E}[\text{ret}_{it-1}]. \qquad (3) \end{aligned}$$

Table 6 reports the results for all firms. Panel A reports the results for all jumps (regardless of the sign of jumps). The news tone is statistically significantly related to the jump size mean. Interestingly, news count is negatively related to jump size mean and percent of uncertain words is positively related to jump size mean. We explore this in Panel B and Panel C when we split the jumps into positive and negative jumps. As shown in Panel B, the news count is significantly and positively related to positive jump returns – more good news is associated with higher positive returns on jump days. In Panel C, the news count is significantly and negatively related to negative jump returns – more bad news is associated with more negative returns on jump days. When we pull all the jump returns in the same regressions, the effect from negative jump returns dominate, resulting the negative coefficient of news count in Panel A. The R^2 of news flows and negative jump returns is 2.18% and it is 1.28% for positive jump returns.[12] In addition, the positive association between percent

[12] The results for large, medium, and small firms are similar and presented in the appendix Tables A4-A6.

of uncertain words and positive jump returns is very surprising.

$\boxed{\text{Table 6 about here}}$

We also conduct the same analysis for the top 20 firms and present the results in Table 7. Compared to the results using all firms, there are some similarities and some differences. For example, news tone is positively associated with jump returns but it is mainly driven by negative jump returns. The percent of uncertain words is no longer statistically significant in explaining the jump returns. The result on news count is similar to the full sample: news count negatively relates to jump returns and is mainly due to the negative jump returns. The R^2 of news flows explaining negative jump returns is close to 26%, compared to an R^2 of around 7% for positive jump returns.

$\boxed{\text{Table 7 about here}}$

For the jump size volatility, we use the same set of variables as in probability of jump analysis and run the following regressions

$$\log(r_{it}^2|\text{Jump}) = c_0 + c_1 \times \text{NewsCount}_{it} + c_2 \times |\text{NewsTone}_{it}| + c_3 \times \text{UncWords}_{it} + c_4 \times |\text{ret}_{it-1}| + \epsilon_{it}, \quad (4)$$

where $r_{it}^2|\text{Jump}$ measures the realized jump return variance. The regression also implies the following relation between the jump size second moment and news

$$\begin{aligned} \mathbb{E}[\log(r_{it}^2|\text{Jump=1})] &= c_0 + c_1 \times \mathbb{E}[\text{NewsCount}_{it}] + c_2 \times \mathbb{E}[|\text{NewsTone}_{it}|] \\ &\quad + c_3 \times \mathbb{E}[\text{UncWords}_{it}] + c_4 \times \mathbb{E}[|\text{ret}_{it-1}|]. \qquad (5) \end{aligned}$$

The results are reported in Table 8. Panel A (B) presents the results for all firms (top 20 firms). In general, the news count and the absolute value of news tone are positively and significantly related to the jump size volatility. Surprisingly, the percent of uncertain words is negatively related to jump size volatility. The R^2 of the jump size volatility regressions are also much higher compared to the jump size mean regressions. For example, in the case of $J99$, the R^2 is 12% for all firms and 24% for top 20 firms. The results are similar for the three size group firms as shown in appendix Table A.7. Overall, our results indicate that news flow is very important in explaining variations in the second moment of jump size distribution.

Table 8 about here

3.3 Determinants of Sensitivity of Jump Probability to News

The results so far suggest that jumps in individual stock returns are significantly related to news counts. We now further explore the potential determinants of this sensitivity of jump probability to news counts. Specifically, we first run logistic regressions, similar to (1), at the firm-by-firm level to obtain an estimate of the sensitivity, $b_{i,1}$, of jump probability to news count for each firm i. Then, we run a cross-sectional regression to understand which firm characteristics are linked to that coefficient. We consider firm characteristics similar to those in Fang and Peress (2009); computed as an annual average of end-of-year observations during the same sample period. The size variable is measured by market capitalization reported by CRSP, book value is taken from Compustat, analyst coverage and dispersion data are collected from I/B/E/S summary files, and the fraction of institutional ownership is from aggregate Thompson Reuters 13F filings. We use the natural logarithm of one plus the number of analysts covering the company for the analyst coverage variable; and one minus the fraction of institutional ownership to represent individual ownership.

Table 9 reports cross-sectional regression results in which the dependent variable $b_{i,1}$ is the coefficient from firm-by-firm logistic regressions estimating the sensitivity of jump probability to news counts for each firm, as explained in the previous paragraph. The results indicate that, from our set of firm characteristics, analyst coverage is the most significant variable in determining the sensitivity of jump probability to news. This relationship is positive, suggesting that firms covered by more analysts are more likely to jump when public news arrives. Fang and Peress (2009) show that analyst and media coverage are substitutes rather than complements. With this in mind, our findings suggest that when public news arrives for firms with higher analyst coverage, the information contained in the news may be more informative, possibly not previously covered by the analysts, thus creating a higher chance of inducing jumps in stock returns.

The fraction of individual ownership shows a significantly negative relationship with respect to the sensitivity of the jump probability to news. In other words, firms with higher institutional ownership tend to jump more frequently when public news arrives. This finding suggests that it is more likely that institutional traders are the ones who trade when public news arrives, at least on the same day. Individual investors may suffer from limited attention and may not be able to react on the same day for all public news that arrives.

Overall, the results of Table 9 suggest that the existence of more information intermediaries, such as analysts, or a higher fraction of institutional investors, helps to incorporate the firm-specific public news into stock return jumps.

Table 9 about here

4 Modelling Time-Series of News and Jumps

Motivated by the findings of previous section, we now move on to the parametric modelling of the time-series of daily stock returns and firm-specific news arrivals. We build on the baseline GARJI model in Maheu and McCurdy (2004) by embedding the observed news arrivals in the latent jump intensity dynamics.

4.1 Dynamics of Continuously Compounded Returns

We define the continuously compounded excess return on an individual firm i as

$$r_t \equiv r_{i,t} - r_{f,t}, \tag{6}$$

in which $r_{i,t}$ is the continuously compounded return (including distributions) on firm i and $r_{f,t}$ is the continuously compounded risk-free rate. Henceforth, we usually refer to r_t, the excess continuously compounded return on firm i, as the log return.

Assume that the dynamics of realized log returns are driven by

$$r_{t+1} = \mu_t + \epsilon_{t+1}, \tag{7}$$

where

$$\begin{aligned} \epsilon_{t+1} &= \epsilon_{1,t+1} + \epsilon_{2,t+1}, && (8) \\ \epsilon_{1,t+1} | \Phi_t &\sim \mathrm{N}(0, \sigma_t^2), && (9) \\ \epsilon_{2,t+1} &= \sum_{k=1}^{n_{t+1}} Y_{t+1,k} - \theta \lambda_{t+1}, && (10) \end{aligned}$$

$\epsilon_{1,t+1}$ is a mean-zero normal innovation to returns directed by a conditional normal process; $\epsilon_{2,t+1}$ is a jump innovation to returns, compensated so that it is mean zero. The jump component $\epsilon_{2,t+1}$ follows a compensated compound Poisson process with intensity λ_{t+1} and individual jump size $Y_{t+1,k}$ drawn from an i.i.d. normal distribution with mean θ and variance δ^2. The innovations $\epsilon_{1,t+1}$ and $\epsilon_{2,t+1}$ are assumed to be *contemporaneously* independent.

The conditional expected return is captured by μ_t. To best understand the impact of news articles on jump dynamics, to begin with we focus on modelling the return innovation dynamics and simplify the asset pricing side of the analysis. In this case, we assume μ_t to be a constant.

First, the diffusive volatility $\sigma_{i,t}$ follows the GARCH process below

$$\sigma_{i,t+1}^2 = \omega + G\epsilon_t^2 + \beta_1\sigma_{i,t}^2 \tag{11}$$

The exact functional form of G coefficient can be flexible. We follow Maheu, McCurdy, and Zhao (2013) and define G as follows.

$$G = \exp(\alpha + I(\epsilon_t < 0)(\alpha_a + \alpha_{a,j}E[n_t|\Phi_t]) \tag{12}$$

Before we specify the dynamics of the latent jump intensity process for the expected number of jumps, λ_{t+1}, we first define the jump intensity residuals ξ_{t+1} as:

$$\xi_{t+1} = E[n_{t+1}|\Phi_{t+1}] - \lambda_{t+1} \tag{13}$$

Since we do not observe the number of jumps directly, we need to use an analytical filtering method to compute $E[n_{t+1}|\Phi_{t+1}]$. Using the jump intensity residual, the dynamics of the jump arrival process can then modeled, as in Chan and Maheu (2002), as a latent autoregressive process labelled GARJI:

$$\lambda_{t+1} = \gamma_0 + \gamma_1 \lambda_t + \gamma_2 \xi_t \tag{14}$$

In contrast to a latent autoregressive process for jump intensity dynamics, we now introduce the new model, which we label as the GJI-N model, by embedding the observable firm-specific news arrival in the dynamics of λ_{t+1} as follows:

$$\lambda_{t+1} \quad = \quad \gamma_2 \xi_t + \phi \lambda_{t+1}^N \tag{15}$$

λ_{t+1}^N, which is measurable with respect to the information set up to time t, utilizes the observable firm-specific news arrival we studied in the previous section. The critical difference, relative to the GARJI model, is that the jump intensity process includes λ_{t+1}^N. More details on the time-series construction of λ_{t+1}^N follows in the next subsection.

Our specification of the GJI-N model only differs from the GARJI model in terms of the specification of latent jump intensity dynamics λ_{t+1}. Therefore, a comparison of the empirical fit of the two models will allow us to study how successfully we can capture the jump intensity dynamics using the observed firm-specific news.

4.2 Filtering the News Arrival Intensity

In the specification of the GJI-N model above, the firm-specific news arrival intensity process λ_{t+1}^N is the critical component that distinguishes our model from the existing GARCH-class of models as it incorporates the observable co-variates, namely the daily count of firm-specific news. We need to make a parametric assumption in order to filter the high degree of noise associated with the daily news count. To do so, we employ the standard Kalman Filter to estimate parameters and filter the observed news count assuming that it follows a discrete mean-reverting process. We assume that a simple mean-reverting process is sufficient in capturing the essential characteristics of the observed news count time-series without complicating the problem too much. The one-dimensional state transition equation is thus given below where N_t is the latent state variable associated with the news arrival intensity.

$$N_{t+1} = N_t + a(\bar{N} - N_t) + \psi e_t$$

Parameter a measures the speed of mean-reversion, $\bar{N}$ represents the long-run mean, and ξ represents the standard deviation of noise term as we assume e_t is a standard normal noise term.

Next, we assume the following simple measurement equation where the news count is observed with normally distributed measurement error $\Lambda e_t^m \sim N(0, \Lambda^2)$.

$$\text{NewsCount}_t = N_t + \Lambda e_t^m$$

The estimated parameters using the likelihood metric from Kalman Filter are reported in Table 10. To visually see the difference between the raw (NewsCount_t) and filtered (N_t) news counts, Figure 1 plots the two time-series for 4 selected companies. We observe that filtered news count in red successfully captures the trend associated with the news count by removing large noise associated with the raw news count in blue.

Table 10 about here

Figure 1 about here

Given the estimated parameters, we now define the conditional expected news flow as follows

$$\lambda_{t+1}^{N} = E_t[N_{t+1}] = \hat{N}_t + a(\bar{N} - \hat{N}_t)$$

where $\hat{N}_t$ denotes the filtered (updated) state variable at time t.

4.3 Estimation

Estimation of the model follows the standard maximum likelihood estimation where the probability of observing daily return residual ϵ_{t+1} is given by:

$$P(\epsilon_{t+1}|\Phi_t) = \sum_{j=0}^{\infty} f(\epsilon_{t+1}|n_{t+1} = j, \Phi_t)P(n_{t+1} = j|\Phi_t)$$

The details regarding the construction of each term on the right-hand side follows Maheu and McCurdy (2004) and Maheu, McCurdy, and Zhao (2013) where all terms have closed-form expression making the estimation straightforward.

5 Estimation Results

We use the daily returns obtained from CRSP (Center for Research in Security Prices) for the same 20 selected firms as in the non-parametric analysis. The in-sample period is from January 2000 to December 2009. As described in the previous section, we first run standard Kalman Filter on the observed daily news count to filter the news arrival intensity, then sequentially run maximum likelihood estimation on the daily returns with news arrival intensity fixed.

Table 11 provides the parameter estimates associated with the jump specifications along with t-stats in parentheses. To save space, the parameter estimates associated with the diffusive volatility specifications are reported in Table A.8 of the Appendix. We also estimate the benchmark GARJI model for comparison purposes and report the parameter estimates associated with the GARJI model in Table A.10. Lastly, to compare the empirical fit with the benchmark GARJI model that does not use observed news count, we provide the log-likelihood of the GARJI model estimates in the last column.

We observe a few interesting findings in Table 11. Most importantly, the ϕ parameter that measures the impact of firm-specific news arrival intensity on the latent jump intensity is positive for all 20 firms and statistically significant at 1% level for most of the firms with an exception for UTC. This is consistent with the non-parametric analysis in the previous sections, that is, that the jump intensity increases with more news arriving. Next, the log-likelihood of the GJI-N model is in general larger than that of the benchmark GARJI model. 18 out of 20 firms have a larger log-likelihood for the GJI-N model as compared to the GARJI model; while 2 firms have marginally smaller log-likelihood. Note that the GJI-N model does not nest the GARJI model as a special case. It is not obvious whether the GJI-N model should provide a better empirical fit than the GARJI model since the GJI-N model has much stricter restrictions in that it has to fit the latent jump intensity dynamics using observed news count variables. In other words, the GJI-N model will provide a better empirical fit than the GARJI model only if the news count variable is strong enough to anchor the news intensity dynamics and identify that dynamic better than a completely latent autoregressive process. Lastly, the estimates of the jump size parameter θ is very close to 0 on average and mostly not statistically significant. This is in line with the previous findings that individual firms experience a similar magnitude of both positive and negative jumps (e.g. Maheu and McCurdy (2004)).

We conclude that the GJI-N model performs at least as well as the GARJI model in explaining the daily return dynamics for the sample of 20 firms analyzed in this section. Note that the only structural difference between two models is the specification of the jump intensity dynamics so we were not expecting a dramatic increase in the log-likelihood. Rather the results serve as evidence, consistent with the previous non-parametric result, that firm-specific news arrival is a good explanatory variable for the latent jump intensity of individual firms.

Table 11 about here

6 Robustness and Additional Analysis

In this section, we summarize some investigations of the robustness of our results. We also present results from additional nonparametric analyses related to our main results concerning the the relationship between measures of news flow and stock return jumps.

6.1 Robustness Analysis using Novel News

In our benchmark analysis we include all news articles. The implicit assumption is that the total news count (including the repeated ones) would capture the importance of news. Nonetheless, it is important to investigate how novel (innovative or surprising) news is related to stock market jumps. That is, does using only novel news as the measure of information flow result in a different relationship with stock market jumps?

To implement this analysis, we rely on the RavenPack news dataset. The RavenPack news dataset provides a variable that measures how "novel" a news article is by comparing the content of the news article with previous news article about the same company. The highest novelty score is 100. We keep only the news articles with novelty score of 100 for this analysis to focus on news that is most likely to ba surprise. In addition to the number of novel news, we also measure the tone of these news articles using the proprietary sentiment measure that RavenPack provides.[13]

Using these textual measures related to novel news flow, we repeat our baseline regressions of Table 5 and present the results in Table 12. In general, our results from Table 12 are quite consistent with those in Table 5: more news are related to higher probabilities of jumps in stock returns. The R^2s of the regressions are in general higher using the novel news articles.

To explore this further. we repeat the same regressions using all news articles from the RavenPack dataset and present the results in Table 13. The results with all news articles in RavenPack are very close to the results with novel news articles only. The differences in results between all news articles by RavenPack in Table 13 and all news articles from Factiva, presented in Table 5, can be explained by differences in news coverage. On one hand, RavenPack's coverage is from 2000 to 2012 (we use the Dow Jones version of RavenPack). On the other hand, Factiva's coverage is from 1980 to 2012 and Factiva covers more sources of news articles.[14] To address the sample-period difference, we present the results using Factiva articles between 2000 and 2012 in Table 14. The results in Table 14 are closer to Table 13 than Table 5.

In summary, we can conlude that our results with respect to the effects of the news flow on stock market jumps is robust to whether we measure the news flow using all news articles

[13] The sentiment measure from RavenPack ranges from 0 to 100; we subtract 50 so that it ranges from -50 to 50 with a negative value of the recentred measure representing negative sentiment and a positive value representing positive sentiment.

[14] According to Factiva, Dow Jones is part of the Factiva news sources and there are 32,000 sources in total for Factiva.

or just novel news articles. The twenty extra years of data provided by the Factiva data collected for this paper allows us to evaluate longer time periods.

Table 12, 13, 14 about here

6.2 Additional Analysis using Different News Categories

Another interesting question we can address with the RavenPack dataset is to what extent different types of news affect the probability of jumps in stock returns. We rely on the news categories that RavenPack provides and regroup them into nine major news categories. We repeat the analysis of Table 5, disaggregating all news into the nine categories, and present the results in Table 15. Again, the results using novel news and all news are similar. Further, the most important news category (according to both t-stats and odd ratios) is Analyst Ratings information, followed by: Credit Ratings information; Earnings and Revenues related information; and Capital Structure information. Consistent with the results from Table 5, the absolute tone of different news categories do not play a significant role in most cases. An interesting exception is the Earnings and Revenues news category for which the absolute news tone is more significant and has higher odds ratio than the news count.

Table 15 about here

7 Conclusions

Stock prices exhibit large, discrete movements, typically labelled as "jumps". A potential important source of jumps in stock returns can be material news events, such as earnings surprises. In this paper, we explicitly test the relationship between jumps in stock returns and news flows. Explicitly analyzing the impact of news on return jumps requires firm-level news data. To do so, we collect 21 million news articles associated with more than 9000 publicly-traded companies from the Factiva database and use textual analysis to derive measures summarizing those news, including news frequency, tone and uncertainty.

Our empirical analyses show that the these measures of news flow content are significantly related to nonparametric measures of jump intensity and jump size distributions and explain an important fraction of variations in the jumps across individual companies. Importantly, the nonparametric analyzes provide input for our time-series modelling of firm-level news processes. By modelling the observable news process explicitly and jointly with a latent jump process, we find that news are important drivers of the jumps in stock returns. Consequently, we are able to enrich the economic content of the widely-used econometric models of jumps used for applications such as option pricing and risk management where stock return jump models are frequently used. Our results have broad implications for these applications of stock return jump models to incorporate news flows explicitly.

References

Andersen, Torben G., Luca Benzoni, and Jesper Lund, 2002, An empirical investigation of continuous-time equity return models, *Journal of Finance* 62, 1239–1284.

Antweiler, Werner, and Murray Z. Frank, 2004, Is all that talk just noise? the information content of internet stock message boards, *The Journal of Finance* 59, 1259–1294.

Bates, David S., 2000, Post-'87 crash fears in the s&p 500 futures option market, *Journal of Econometrics* 94, 181–238.

———, and Roger Craine, 1999, Valuing the futures market clearinghouse's default exposure during the 1987 crash, *Journal of Money, Credit, and Banking* 31, 248–272.

Bekaert, Geert, and Stephen F. Gray, 1998, Target zones and exchange rates: An empirical investigation, *Journal of International Economics* 45, 1–35.

Berry, Thomas D, and Keith M Howe, 1994, Public information arrival, *Journal of Finance* 49, 1331–1346.

Bollerslev, Tim, and Viktor Todorov, 2011, Tails, fears and risk premia, *Journal of Finance* 66, 2165–2211.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson, 2018, Information, Trading, and Volatility: Evidence from Firm-Specific News, *The Review of Financial Studies* 32, 992–1033.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu, 2019, The structure of economic news, *Working paper*.

Chan, Wing Hong, and John M. Maheu, 2002, Conditional jump dynamics in stock market returns, *Journal of Business & Economic Statistics* 20, 377–389.

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

Chernov, Mikhail, Ronald A. Gallant, Eric Ghysels, and George Tauchen, 2003, Alternative models for stock price dynamics, *Journal of Econometrics* 116, 225–257.

Christoffersen, Peter, Kris Jacobs, and Chayawat Ornthanalai, 2012, Dynamic jump intensities and risk premiums: Evidence from s&p500 returns and options, *Journal of Financial Economics* 106, 447 – 472.

Cutler, David M., James M. Poterba, and Lawrence H. Summers, 1989, What moves stock prices?, *The Journal of Portfolio Management* 15, 4–12.

Das, Sanjiv R., 2002, The surprise element: Jumps in interest rate, *Journal of Econometrics* 106, 27–65.

Engle, Robert F., Martin Klint Hansen, and Asger Lunde, 2012, And now, the rest of the news: Volatility and firm specific news arrival, *Working Paper.*

Eraker, Bjorn, Michael S. Johannes, and Nicholas G. Polson, 2003, The impact of jumps in volatility and returns, *Journal of Finance* 58, 1269–1300.

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

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy, 2018, Text as data, *Journal of Economic Literature, Forthcoming.*

Gilder, Dudley, Mark B. Shackleton, and Stephen J. Taylor, 2014, Cojumps in stock prices: Empirical evidence, *Journal of Banking & Finance* 40, 443 – 459.

Johannes, Michael, Rohit Kumar, and Nicholas G. Polson, 1999, State dependent jump models: How do U.S. equity indices jump?, Working Paper, Graduate School of Business, University of Chicago.

Ke, Zheng Tracy, Bryan T. Kelly, and Dacheng Xiu, 2019, Predicting returns with text data, *Working paper.*

Kelly, Bryan T., Asaf Manela, and Alan Moreira, 2019, Text selection, *Working paper.*

Kyle, Albert S., Anna A. Obizhaeva, Nitish Ranjan Sinha, and Tugkan Tuzun, 2011, News arrival and trading game invariance, Working Paper.

Lee, Suzanne S., 2012, Jumps and information flow in financial markets, *Review of Financial Studies* 25, 440–479.

———, and Per A. Mykland, 2008, Jumps in financial markets: A new nonparametric test and jump dynamics, *Review of Financial Studies* 21, 2535–2563.

Loughran, Tim, and Bill McDonald, 2011, When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, *The Journal of Finance* 66, 35–65.

———, 2013, IPO first-day returns, offer price revisions, volatility, and Form S-1 language, *Journal of Financial Economics* 109, 307 – 326.

———, 2016, Textual analysis in accounting and finance: A survey, *Journal of Accounting Research* 54, 1187–1230.

Maheu, John M., and Thomas H. McCurdy, 2004, News arrival, jump dynamics, and volatility components for individual stock returns, *Journal of Finance* 59, 755–794.

——— , and Xiaofei Zhao, 2013, Do jumps contribute to the dynamics of the equity premium?, *Journal of Financial Economics* 110, 457 – 477.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137 – 162.

Mitchell, Mark L, and J Harold Mulherin, 1994, The impact of public information on the stock market, *Journal of Finance* 49, 923–950.

Neely, Christopher J., 1999, Target zones and conditional volatilty: The role of realignments, *Journal of Empirical Finance* 6, 177–192.

Pan, Jun, 2002, The jump-risk premia implicit in options: Evidence from an integrated time-series study, *Journal of Financial Economics* 63, 3–50.

Press, S. James, 1967, A compound events model for security prices, *Journal of Business* 40, 317–335.

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

——— , 2010, Does public financial news resolve asymmetric information?, *Review of Financial Studies* 23, 1437–1467.

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

——— , Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More than words: Quantifying language to measure firms' fundamentals, *The Journal of Finance* 63, 1437–1467.

Veldkamp, Laura L., 2006, Media frenzies in markets for financial information, *American Economic Review* 96, 577–601.

Table 1: **Summary Statistics of News Counts**

This table reports the summary statistics of news counts downloaded from the Factiva database. Panel A reports the summary statistics for all firms and firms in three size groups with equal number of firms in each group, sorted by their market capitalizations at the end of the sample period. Panel B reports the summary statistics for the top 20 firms with the most news counts. In each panel, the first column reports the total number of news articles found for each firm or size group during the sample period. The second to fourth columns report daily mean, median, and standard deviation of news counts for each firm or size group. The last column reports the percentage of news post 2000. The sample period is from January 1980 to July 2012.

	Summary Statistics				
Company	Total	Mean	Median	Std. Dev.	% Post-2000
			Panel A: All Firms		
All Firms	21,510,023	1.06	0	10.45	79.48%
Large Firms	18,490,700	1.95	0	14.88	80.44%
Medium Firms	2,208,347	0.37	0	3.86	77.71%
Small Firms	810,976	0.17	0	1.36	62.29%
			Panel B: Top 20 News Firms		
Amazon	63,092	16.48	12	15.51	89.69%
American Express	78,725	9.58	5	20.39	65.59%
AT&T	149,261	20.80	7	52.58	87.75%
Bank of America	209,106	25.45	4	74.98	91.56%
Chevron	123,636	15.24	7	18.84	80.65%
Cisco	123,074	21.76	15	81.86	88.42%
Disney	180,787	21.99	9	35.36	84.73%
Ebay	72,096	20.69	17	17.55	97.12%
GE	308,872	37.58	16	64.80	82.16%
IBM	258,977	31.52	22	58.15	60.80%
Intel	196,668	23.95	9	56.53	83.14%
Johnson & Johnson	108,214	13.16	4	21.07	89.58%
JP Morgan	242,184	29.48	5	68.09	92.51%
Merck & Co.	52,237	6.35	2	10.96	88.82%
Microsoft	388,507	58.40	48	60.43	80.61%
Pfizer	119,741	14.58	4	21.84	86.61%
UTC	66,258	8.06	4	13.87	75.13%
Verizon	158,642	22.11	4	66.82	94.87%
Wal Mart	170,208	20.71	4	51.13	92.58%
Yahoo	88,592	21.59	15	26.55	90.93%
Top 20 News Firms	3,158,877	21.82	8	49.76	84.20%

Table 2: **Summary Statistics of Daily News Tones**

This table reports the summary statistics of daily news tones and percent of uncertain words (times 1,000). Daily new tone variable is constructed by analyzing the first paragraph of each news article. We calculate the percentage of positive and negative words using the list from Loughran and McDonald (2011). Then, tone of individual articles are aggregated at the daily level using the total number of words in each article as weights. The sample period is from January 1980 to July 2012.

Company	Summary Statistics			
	Average Tone	Std. Dev. Tone	Average Uncertain Words	Std. Dev. Uncertain Words
	Panel A: All Firms			
All Firms	-0.315	17.122	2.185	14.661
Large Firms	-0.506	18.445	2.959	14.817
Medium Firms	0.042	15.074	1.519	13.318
Small Firms	-0.385	16.801	1.480	15.844
	Panel B: Top 20 News Firms			
Amazon	-0.946	20.436	3.228	6.602
American Express	-6.734	26.318	6.886	15.499
AT&T	-5.424	21.349	5.499	15.888
Bank of America	-10.327	25.303	7.753	19.843
Chevron	-9.227	22.902	6.312	14.690
Cisco	2.086	18.295	2.266	8.679
Disney	-3.462	20.094	4.739	13.735
Ebay	-4.924	20.999	3.133	7.623
GE	-7.010	19.237	3.875	10.447
IBM	-1.652	19.392	3.098	10.086
Intel	-3.213	21.326	4.469	14.486
Johnson & Johnson	-5.354	25.427	6.617	19.704
JP Morgan	-8.721	23.142	6.500	17.529
Merck & Co.	-2.948	25.524	5.709	19.599
Microsoft	-5.403	20.074	2.858	11.223
Pfizer	-6.188	23.931	5.819	16.879
UTC	-6.850	25.267	7.362	18.066
Verizon	-5.471	20.906	5.594	15.677
Wal Mart	-6.418	21.367	4.719	15.374
Yahoo	-1.569	19.426	2.848	8.328
Top 20 News Firms	-5.321	22.604	5.198	15.245

Table 3: **Summary Statistics of Days with Realized Jumps**

This table reports summary statistics of daily realized jumps. The daily return jump indicator is identified using 4 different statistics. $J99$ and $J95$ indicator uses Lee and Mykland (2008)'s Lemma 1 statistic at 99% and 95% significance, respectively. $J_0 99$ and $J_0 95$ indicator uses a less tight bound from the normal distribution as in the Theorem 1 of Lee and Mykland (2008). We also use the correction term in Gilder, Shackleton, and Taylor (2014). Each of four statistics $\{J99, J95, J_0 99, J_0 95\}$ thus identifies the jump day if the absolute value of daily return is above $\{5.1024, 4.4881, 3.2283, 2.4565\}$ times the daily spot volatility. The sample period is from January 1980 to July 2012.

Company	N	Jump Days $J99$	Jump Days $J95$	Jump Days $J_0 99$	Jump Days $J_0 95$
		Panel A: All Firms			
All Firms	20,079,694	351,374	480,489	1,096,893	2,055,081
Large Firms	9,426,014	117,683	172,332	445,746	894,227
Medium Firms	5,950,623	111,943	152,270	341,381	627,889
Small Firms	4,703,057	121,748	155,887	309,766	532,965
		Panel B: Top 20 News Firms			
Amazon	3,811	53	70	159	306
American Express	8,203	49	89	285	674
AT&T	7,158	43	79	247	604
Bank of America	8,200	63	97	308	692
Chevron	8,097	32	61	278	673
Cisco	5,638	38	63	188	441
Disney	8,204	59	96	313	670
Ebay	3,468	33	51	128	270
GE	8,203	37	64	267	681
IBM	8,199	63	101	283	654
Intel	8,194	49	86	275	652
Johnson & Johnson	8,204	50	85	304	691
JP Morgan	8,199	57	99	298	708
Merck & Co.	8,203	57	93	293	696
Microsoft	6,635	63	91	231	559
Pfizer	8,194	50	79	273	633
UTC	8,204	49	91	282	662
Verizon	7,159	36	66	242	586
Wal Mart	8,203	44	88	281	666
Yahoo	4,087	43	56	172	348
Top 20 News Firms	144,463	968	1,605	5,107	11,866

Table 4: **Effect of News Counts on Probability of Daily Jump (All Firms)**

This table reports coefficients from the pooled logistic regression of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for all firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panel B reports the odds ratios of each variable in brackets. All regression specifications include a constant term that is not reported for brevity.

	(1)	(2)	(3)	(4)
	$J99$	$J95$	$J_0 99$	$J_0 95$
	Panel A: Coefficient Estimates			
$NewsCount_t$	0.2004	0.1915	0.1598	0.1318
	(105.47)	(112.65)	(114.14)	(109.83)
$\|NewsTone_t\|$	0.0186	0.0156	0.0113	0.008
	(10.94)	(10.40)	(10.27)	(8.89)
$UncWords_t$	0.0094	0.0093	0.0049	0.0037
	(5.53)	(6.20)	(4.45)	(4.11)
$\|Ret_{t-1}\|$	0.0809	0.0897	0.1016	0.1061
	(44.94)	(59.80)	(92.36)	(117.89)
N	20,079,694	20,079,694	20,079,694	20,079,694
$R^2_{McFadden}$	1.40%	1.24%	0.83%	0.61%
	Panel B: Odds Ratios			
$NewsCount_t$	[1.222]	[1.211]	[1.173]	[1.141]
$\|NewsTone_t\|$	[1.019]	[1.016]	[1.011]	[1.008]
$UncWrods_t$	[1.009]	[1.009]	[1.005]	[1.004]
$\|Ret_{t-1}\|$	[1.084]	[1.094]	[1.107]	[1.112]

Table 5: **Effect of News Counts on Probability of Daily Jump (Top 20 Firms)**
This table reports coefficients from logistic regressions of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for the top 20 large firms. The first row reports the result from the pooled logistic regression for the top 20 news firms and the rest of the table reports logistic regression result for each individual firm. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. The odds ratios are reported in brackets.

	$NewsCount_t$	$\lvert NewsTone_t \rvert$	$UncWords_t$	$\lvert Ret_{t-1} \rvert$	$R^2_{McFadden}$
			Top 20 Firms Total		
Total	0.1200	0.1577	-0.1022	0.0323	0.85%
	(5.02)	(5.59)	(-3.01)	(1.38)	
	[1.127]	[1.171]	[0.903]	[1.033]	
			Individual 20 Firms		
Amazon	0.048	19.862	-56.227	-1.208	16.51%
	(7.57)	(3.57)	(-1.47)	(-0.35)	
American Express	0.005	-0.520	-10.650	1.517	0.59%
	(2.53)	(-0.07)	(-0.89)	(0.21)	
AT&T	0.001	-3.670	-0.913	15.783	0.69%
	(1.46)	(-0.58)	(-0.12)	(2.34)	
Bank of America	0.001	0.483	4.162	3.909	0.28%
	(1.37)	(0.08)	(0.68)	(1.13)	
Chevron	0.007	12.097	-4.271	15.613	1.13%
	(1.21)	(1.73)	(-0.49)	(1.62)	
Cisco	0.001	16.284	-6.587	-0.571	0.96%
	(2.59)	(3.11)	(-0.56)	(-0.08)	
Disney	0.002	2.238	4.580	2.265	0.28%
	(2.21)	(0.32)	(0.66)	(0.39)	
Ebay	0.035	15.784	-110.100	2.236	12.38%
	(6.87)	(1.76)	(-1.05)	(0.58)	
GE	0.002	5.397	9.639	0.961	0.77%
	(3.91)	(0.82)	(1.08)	(0.08)	
IBM	0.002	6.977	-14.523	-0.737	0.49%
	(4.11)	(0.90)	(-0.91)	(-0.08)	
Intel	0.002	20.600	-16.881	-3.869	1.71%
	(3.70)	(4.46)	(-1.84)	(-0.63)	
Johnson & Johnson	0.017	10.760	-11.132	13.005	3.25%
	(5.49)	(1.68)	(-1.21)	(1.49)	
JP Morgan	0.001	15.215	-11.604	1.790	0.95%
	(1.60)	(2.94)	(-1.52)	(0.28)	
Merck & Co.	0.023	-0.609	0.778	-10.506	1.67%
	(3.12)	(-0.12)	(0.16)	(-1.10)	
Microsoft	0.003	12.480	-8.387	-1.856	0.90%
	(2.43)	(2.01)	(-0.90)	(-0.25)	
Pfizer	0.015	5.698	-5.213	-20.841	3.27%
	(5.48)	(0.89)	(-0.72)	(-2.04)	
UTC	0.008	6.079	-3.971	2.604	0.52%
	(3.29)	(0.93)	(-0.50)	(0.34)	
Verizon	0.001	2.202	5.675	11.084	0.48%
	(3.17)	(0.26)	(0.57)	(0.88)	
Wal Mart	0.001	8.441	-9.146	7.248	0.51%
	(2.37)	(1.64)	(-0.97)	(0.89)	
Yahoo	0.017	18.052	-199.800	-8.498	11.14%
	(6.03)	(1.92)	(-1.95)	(-1.39)	

Table 6: **Effect of News Counts on Daily Jump Size (All Firms). 1980-2012**
This table reports coefficients from regressions of daily jump sizes conditional on the jump indicator being 1 on daily news count and news tone for all firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panels B and C report the results for positive jump sizes and negative jump sizes, respectively. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: All Jumps			
$NewsCount_t$	-2.45E-04	-2.38E-04	-1.51E-04	-2.04E-05
	(-7.81)	(-9.19)	(-10.77)	(-5.31)
$NewsTone_t$	1.0491	0.8733	0.5197	0.3472
	(45.02)	(47.78)	(54.08)	(58.97)
$UncWords_t$	0.6284	0.5050	0.2611	0.1623
	(23.96)	(24.51)	(23.53)	(23.72)
Ret_{t-1}	-0.2096	-0.2206	-0.2586	-0.2848
	(-48.14)	(-63.01)	(-130.22)	(-220.97)
N	351,374	480,489	1,096,893	2,055,021
R^2	1.29%	1.34%	1.81%	2.49%
	Panel B: Positive Jumps			
$NewsCount_t$	6.09E-04	3.85E-04	4.95E-05	-7.42E-05
	(11.19)	(9.16)	(2.34)	(-5.77)
$NewsTone_t$	0.3343	0.2559	0.1486	0.0967
	(11.31)	(11.28)	(12.78)	(13.60)
$UncWords_t$	0.5382	0.4234	0.2524	0.1580
	(16.01)	(16.42)	(18.60)	(18.89)
Ret_{t-1}	-0.2522	-0.2464	-0.2301	-0.2183
	(-48.25)	(-58.92)	(-97.06)	(-139.46)
N	206,070	282,470	629,554	1,140,445
R^2	1.28%	1.32%	1.52%	1.70%
	Panel C: Negative Jumps			
$NewsCount_t$	-4.55E-04	-4.09E-04	-1.92E-04	-8.82E-06
	(-26.77)	(-27.75)	(-22.18)	(-4.34)
$NewsTone_t$	0.4710	0.4012	0.2424	0.1688
	(28.81)	(30.35)	(33.30)	(36.93)
$UncWords_t$	0.0225	0.0117	-0.0472	-0.0339
	(1.24)	(0.79)	(-5.69)	(-6.47)
Ret_{t-1}	-0.0412	-0.0589	-0.0968	-0.1252
	(-12.52)	(-22.20)	(-63.34)	(-125.79)
N	145,304	198,019	467,339	914,636
R^2	2.18%	1.89%	1.63%	2.07%

Table 7: **Effect of News Counts on Daily Jump Size (Top 20 Firms). 1980-2012**

This table reports coefficients from regressions of daily jump sizes conditional on the jump indicator being 1 on daily news count and news tone for the top 20 news firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panels B and C report the results for positive jump sizes and negative jump sizes, respectively. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
		Panel A: All Jumps		
$NewsCount_t$	-1.49E-04	-1.42E-04	-5.30E-05	-3.47E-05
	(-3.58)	(-4.37)	(-3.14)	(-3.96)
$NewsTone_t$	0.8844	0.7251	0.4643	0.3334
	(6.35)	(7.46)	(10.65)	(13.66)
$UncWords_t$	0.2050	0.3308	0.3339	0.2572
	(0.72)	(1.88)	(5.13)	(6.76)
Ret_{t-1}	0.2604	0.2600	0.1441	0.1368
	(2.54)	(3.46)	(4.33)	(6.96)
N	968	1,605	5,107	11,866
R^2	7.05%	6.09%	2.98%	2.30%
		Panel B: Positive Jumps		
$NewsCount_t$	2.26E-04	1.90E-04	1.13E-04	7.85E-05
	(4.74)	(5.54)	(7.52)	(8.69)
$NewsTone_t$	-0.0877	-0.0751	0.0220	0.0085
	(-0.69)	(-0.91)	(0.60)	(0.41)
$UncWords_t$	-0.3590	-0.2884	-0.0278	-0.0201
	(-1.54)	(-2.14)	(-0.55)	(-0.67)
Ret_{t-1}	-0.2833	-0.2519	-0.1352	-0.0633
	(-3.79)	(-4.45)	(-5.35)	(-4.09)
N	561	949	2,997	6,811
R^2	6.99%	6.12%	2.91%	1.46%
		Panel C: Negative Jumps		
$NewsCount_t$	-1.64E-04	-1.70E-04	-1.37E-04	-5.90E-05
	(-6.42)	(-7.81)	(-10.41)	(-9.61)
$NewsTone_t$	0.4179	0.3979	0.2838	0.2439
	(4.24)	(5.17)	(7.78)	(11.79)
$UncWords_t$	0.0270	0.1361	0.1604	0.1911
	(0.12)	(0.86)	(2.56)	(5.42)
Ret_{t-1}	0.7912	0.6046	0.2627	0.1551
	(8.09)	(8.75)	(7.96)	(8.52)
N	407	656	2,110	5,055
R^2	25.91%	22.33%	11.41%	6.65%

Table 8: **Effect of News Counts on Daily Jump Vol. 1980-2012**

This table reports coefficients from the regression of daily jump volatilities, defined as log of squared jump size conditional on the jump indicator being 1, on daily news count and news tone for all firms and for the top 20 large firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same the mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. All regression specifications include a constant term that is not reported for brevity.

	(1)	(2)	(3)	(4)
	$J99$	$J95$	$J_0 99$	$J_0 95$
	Panel A: All Firms			
$NewsCount_t$	0.0063	0.0050	0.0012	-0.0003
	(24.33)	(22.33)	(8.03)	(-6.27)
$\|NewsTone_t\|$	9.7357	8.0329	5.1597	3.3551
	(46.90)	(47.02)	(48.16)	(43.94)
$UncWords_t$	-3.8959	-2.9717	-1.4210	-0.8314
	(-17.14)	(-15.87)	(-11.82)	(-9.61)
$\|Ret_{t-1}\|$	8.4670	8.7873	9.9479	11.0143
	(213.82)	(260.94)	(429.35)	(619.19)
N	351,374	480,489	1,096,893	2,055,081
R^2	12.46%	13.08%	14.66%	15.84%
	Panel B: Top 20 Firms			
$NewsCount_t$	0.0035	0.0035	0.0024	0.0009
	(7.11)	(8.28)	(8.73)	(5.03)
$\|NewsTone_t\|$	5.4143	5.6891	3.7057	3.9109
	(2.68)	(3.76)	(4.33)	(6.77)
$UncWords_t$	-3.7022	-5.5402	-2.1117	-2.7165
	(-1.11)	(-2.43)	(-1.96)	(-3.59)
$\|Ret_{t-1}\|$	22.6351	23.4320	22.9407	25.3301
	(15.06)	(19.21)	(33.85)	(51.24)
N	968	1,605	5,107	11,866
R^2	23.69%	23.28%	20.20%	19.09%

Table 9: **Cross-sectional Determinants of Sensitivity of Jump Probability to News**

This table reports cross-sectional regression results concerning potential firm characteristic determinants of the sensitivity of jump probabilities to news counts. The dependent variable in the cross-sectional regressions is $b_{i,1}$ estimated from firm-by-firm logistic regressions of Equation 1, that is, the sensitivity of jump probability to news counts at the individual firm level. The firm characteristic regressors in the cross-sectional regressions are computed as an annual average of end-of-year observations during the sample period from January 1980 to July 2012. t-stats are reported in the parentheses.

	(1) $b_{i,1}(J99)$	(2) $b_{i,1}(J95)$	(3) $b_{i,1}(J_0 99)$	(4) $b_{i,1}(J_0 95)$
Intercept	0.3119	0.1655	-0.1401	-0.1539
	(1.09)	(0.59)	(-0.74)	(-0.84)
Size	-0.0419	-0.0240	0.0123	0.0137
	(-1.58)	(-0.93)	(0.71)	(0.81)
Book-to-Market Ratio	-0.0966	-0.0611	0.0283	0.0292
	(-0.85)	(-0.55)	(0.38)	(0.40)
Analyst Coverage	0.2204	0.1495	0.0531	0.0565
	(3.72)	(2.59)	(1.36)	(1.50)
Analyst Dispersion	0.0388	0.0384	0.0215	0.0104
	(1.18)	(1.19)	(0.99)	(0.50)
Individual Ownership	-0.1033	-0.1097	-0.0673	-0.0197
	(-1.96)	(-2.14)	(-1.95)	(-0.59)
Return Volatility	-0.0869	-0.0471	0.0198	0.0444
	(-1.38)	(-0.77)	(0.48)	(1.11)
N	5,440	5,440	5,440	5,440
R^2	0.79%	0.54%	0.48%	0.30%

Table 10: **Parameter Estimates from the first-stage Kalman Filter**

This table reports estimated parameters from the first-stage Kalman filter applied on the daily news count variables for the top 20 large firms. Standard maximum likelihood estimation was used to estimate the parameters. The sample period is from January 2000 to December 2009.

Company Name	a	$\bar{N}$	ψ
Amazon	0.1574	15.12	4.05
American Express	0.0063	17.24	0.89
AT&T	0.0176	37.87	2.79
Bank of America	0.0109	48.03	4.60
Chevron	0.0327	31.02	3.00
Cisco	0.1424	32.38	6.17
Disney	0.0130	49.55	2.37
Ebay	0.0180	22.36	1.84
GE	0.0064	79.30	2.68
IBM	0.0153	51.49	2.57
Intel	0.1664	49.36	11.08
Johnson & Johnson	0.0031	30.09	1.12
JP Morgan	0.0147	63.70	4.80
Merck & Co.	0.0201	13.14	1.08
Microsoft	0.0463	104.48	8.26
Pfizer	0.0066	32.97	1.32
UTC	0.0064	15.17	0.57
Verizon	0.0145	42.39	3.46
Wal Mart	0.0048	49.61	1.63
Yahoo	0.0653	24.39	5.52

Table 11: **Parameter Estimates for the GJI-N Model (Jump Parameters)**

This table reports parameters for the GJI-N model estimated on daily returns and filtered news count. For brevity, we only report the parameters associated with the jump dynamics (intensity and size) and report the rest in the appendix. t-stats computed using the outer product of gradient method are reported in the parentheses. For comparison, we also report the log-likelihood of the benchmark GARJI model etimates in the last column. The sample period is from January 2000 to December 2009.

Company	γ_2	θ	δ	ϕ	lgl	GARJI lgl
Amazon	0.089	0.033	0.113	2.49E-03	5,137.00	5,122.47
	(1.30)	(2.61)	(8.99)	(5.65)		
American Express	0.299	0.002	0.013	1.25E-02	6,371.14	6,370.70
	(1.25)	(1.37)	(7.90)	(2.57)		
AT&T	0.646	-0.003	0.020	1.72E-03	6,757.90	6,758.82
	(1.60)	(-1.12)	(5.43)	(2.32)		
Bank of America	-0.010	-0.007	0.014	1.88E-03	6,683.48	6,664.36
	(-0.05)	(-2.63)	(9.11)	(3.08)		
Chevron	-0.001	-0.003	0.092	2.87E-05	7,019.24	7,010.78
	(0.00)	(-0.04)	(0.26)	(2.83)		
Cisco	0.237	0.012	0.049	1.22E-03	5,859.15	5,853.41
	(1.52)	(1.80)	(7.89)	(3.54)		
Disney	0.070	0.009	0.069	3.59E-04	6,497.85	6,496.25
	(0.65)	(0.79)	(6.59)	(3.04)		
Ebay	0.375	-0.003	0.026	1.73E-03	6,909.07	6,906.96
	(1.65)	(-0.60)	(6.73)	(2.56)		
GE	0.065	0.000	0.034	3.07E-04	6,794.03	6,791.45
	(0.57)	(0.00)	(6.87)	(3.22)		
IBM	0.107	-0.001	0.023	1.52E-03	6,988.98	6,986.74
	(0.99)	(-0.50)	(8.80)	(3.79)		
Intel	-0.050	-0.021	0.066	3.69E-04	5,870.95	5,867.12
	(-0.15)	(-1.42)	(5.82)	(2.76)		
Johnson & Johnson	0.340	0.000	0.013	4.24E-03	7,764.41	7,762.89
	(2.02)	(0.22)	(8.73)	(3.73)		
JP Morgan	0.029	0.005	0.037	4.62E-04	6,265.80	6,254.27
	(0.32)	(0.82)	(7.58)	(3.17)		
Merck & Co.	0.307	0.000	0.061	2.72E-03	6,749.41	6,748.56
	(2.54)	(-0.06)	(13.08)	(5.16)		
Microsoft	0.015	0.001	0.051	3.51E-04	6,626.23	6,627.64
	(0.15)	(0.09)	(10.48)	(4.52)		
Pfizer	0.405	-0.001	0.025	2.71E-03	6,862.14	6,854.59
	(2.03)	(-0.47)	(12.00)	(3.81)		
UTC	0.463	-0.046	0.093	3.26E-04	6,837.66	6,834.48
	(0.57)	(-0.93)	(2.94)	(1.51)		
Verizon	0.386	0.000	0.017	2.78E-03	6,921.38	6,920.74
	(1.48)	(-0.09)	(6.83)	(2.83)		
Wal Mart	0.026	0.003	0.014	3.59E-03	7,037.28	7,027.04
	(0.21)	(1.94)	(7.74)	(2.55)		
Yahoo	0.234	0.017	0.085	1.56E-03	5,263.66	5,250.38
	(1.89)	(2.17)	(11.45)	(4.47)		

Table 12: **Effect of News Counts on Probability of Daily Jump (Top 20 Firms) using Novel News from RavenPack Data**

This table reports coefficients from logistic regressions of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for the top 20 news firms using novel news from RavenPack database. The first row reports the result from the pooled logistic regression for the top 20 large firms and the rest of the table reports logistic regression result for each individual firm. The explanatory variables the total number of news, absolute news tone reported on the Factiva database each day, and absolute value of previous day's return. The sample period is from January 2000 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. The odds ratios are reported in brackets.

	$NewsCount_t$	$\|NewsTone_t\|$	$\|Ret_{t-1}\|$	$R^2_{McFadden}$
	Top 20 Firms Total			
Total	0.350	0.178	-0.053	4.37%
	(10.96)	(2.37)	(-1.70)	
	[1.419]	[1.194]	[0.949]	
	Individual 20 Firms			
Amazon	0.157	0.080	2.184	8.40%
	(4.62)	(6.57)	(0.51)	
American Express	0.225	0.026	2.968	3.59%
	(4.96)	(1.39)	(0.29)	
AT&T	-11.728	-0.014	2.781	0.71%
	(-50.09)	(-0.99)	(0.24)	
Bank of America	0.127	0.043	2.601	3.69%
	(3.70)	(2.14)	(0.62)	
Chevron	0.188	-0.030	10.574	1.34%
	(2.06)	(-1.10)	(0.65)	
Cisco	0.146	0.061	-7.809	4.65%
	(4.94)	(3.23)	(-0.71)	
Disney	0.189	0.029	-15.829	3.11%
	(3.89)	(1.19)	(-1.17)	
Ebay	0.187	0.071	0.307	7.64%
	(5.17)	(5.28)	(0.06)	
GE	0.276	-0.007	1.399	12.34%
	(7.85)	(-0.21)	(0.11)	
IBM	0.165	0.024	-38.914	4.19%
	(5.07)	(0.89)	(-1.84)	
Intel	0.095	0.046	-3.617	2.20%
	(1.73)	(2.34)	(-0.34)	
Johnson & Johnson	0.206	-0.004	-2.321	3.21%
	(4.66)	(-0.18)	(-0.15)	
JP Morgan	0.071	0.033	5.617	2.87%
	(3.88)	(1.80)	(0.80)	
Merck & Co.	0.272	0.037	-13.019	11.50%
	(8.88)	(2.56)	(-0.88)	
Microsoft	0.120	0.011	-8.209	1.22%
	(2.36)	(0.44)	(-0.78)	
Pfizer	0.237	0.053	-18.767	13.53%
	(8.16)	(2.97)	(-1.06)	
UTC	0.291	-0.027	-9.521	7.94%
	(4.89)	(-1.00)	(-0.81)	
Verizon	0.271	0.033	-7.609	8.56%
	(6.02)	(1.51)	(-0.50)	
Wal Mart	0.275	0.005	-2.062	14.42%
	(9.02)	(0.21)	(-0.09)	
Yahoo	0.130	0.051	-6.004	3.56%
	(4.84)	(3.56)	(-0.99)	

Table 13: **Effect of News Counts on Probability of Daily Jump (Top 20 Firms) using All News from RavenPack Data**

This table reports coefficients from logistic regressions of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for the top 20 news firms using all news from RavenPack database. The first row reports the result from the pooled logistic regression for the top 20 large firms and the rest of the table reports logistic regression result for each individual firm. The explanatory variables the total number of news, absolute news tone reported on the Factiva database each day, and absolute value of previous day's return. The sample period is from January 2000 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. The odds ratios are reported in brackets.

	$NewsCount_t$	$\|NewsTone_t\|$	$\|Ret_{t-1}\|$	$R^2_{McFadden}$
	Top 20 Firms Total			
Total	0.342	0.177	-0.042	5.04%
	(14.61)	(2.49)	(-1.33)	
	[1.408]	[1.194]	[0.959]	
	Individual 20 Firms			
Amazon	0.081	0.073	2.947	8.23%
	(6.28)	(6.03)	(0.76)	
American Express	0.105	0.025	3.702	4.14%
	(4.60)	(1.49)	(0.36)	
AT&T	-11.207	-0.033	2.781	0.71%
	(-46.79)	(-2.20)	(0.24)	
Bank of America	0.080	0.032	2.258	6.98%
	(5.10)	(1.41)	(0.52)	
Chevron	0.096	-0.045	10.667	1.86%
	(2.74)	(-1.51)	(0.68)	
Cisco	0.117	0.042	-4.466	8.64%
	(7.88)	(2.36)	(-0.42)	
Disney	0.076	0.033	-15.362	3.34%
	(4.01)	(1.58)	(-1.13)	
Ebay	0.084	0.064	1.089	7.76%
	(6.23)	(4.97)	(0.20)	
GE	0.099	0.002	5.372	11.79%
	(7.54)	(0.06)	(0.45)	
IBM	0.064	0.014	-38.257	3.69%
	(4.66)	(0.66)	(-1.79)	
Intel	0.064	0.048	-2.732	3.82%
	(4.39)	(2.73)	(-0.25)	
Johnson & Johnson	0.068	0.008	-1.338	3.25%
	(4.83)	(0.45)	(-0.09)	
JP Morgan	0.060	0.038	5.332	6.24%
	(6.33)	(1.96)	(0.75)	
Merck & Co.	0.099	0.034	-8.485	8.93%
	(7.14)	(2.60)	(-0.66)	
Microsoft	0.056	0.030	-6.845	2.60%
	(3.95)	(1.28)	(-0.66)	
Pfizer	0.072	0.055	-18.459	10.63%
	(7.19)	(3.39)	(-1.07)	
UTC	0.131	0.003	-8.272	7.99%
	(3.83)	(0.10)	(-0.72)	
Verizon	0.111	0.039	-5.371	7.63%
	(5.53)	(1.80)	(-0.36)	
Wal Mart	0.101	0.005	0.285	13.40%
	(8.82)	(0.22)	(0.01)	
Yahoo	0.071	0.048	-5.041	5.20%
	(6.42)	(3.12)	(-0.84)	

Table 14: **Effect of News Counts on Probability of Daily Jump (Top 20 Firms) using All News from Factiva Data**

This table reports coefficients from logistic regressions of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for the top 20 news firms using Factiva database. The first row reports the result from the pooled logistic regression for the top 20 large firms and the rest of the table reports logistic regression result for each individual firm. The explanatory variables are the total number of news, absolute news tone reported on the Factiva database each day, and absolute value of previous day's return. The sample period is from January 2000 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. The odds ratios are reported in brackets.

	$NewsCount_t$	$\|NewsTone_t\|$	$\|Ret_{t-1}\|$	$R^2_{McFadden}$
		Top 20 Firms Total		
Total	0.235	0.118	-0.054	5.45%
	(4.24)	(3.87)	(-1.56)	
	[1.265]	[1.125]	[0.947]	
		Individual 20 Firms		
Amazon	0.051	13.186	-0.901	17.97%
	(7.58)	(2.11)	(-0.22)	
American Express	0.011	16.524	2.214	2.06%
	(2.29)	(1.71)	(0.20)	
AT&T	0.010	19.160	6.637	2.60%
	(2.84)	(1.13)	(0.60)	
Bank of America	0.001	9.591	2.791	0.93%
	(2.25)	(0.86)	(0.57)	
Chevron	0.016	14.321	11.091	2.42%
	(3.04)	(1.23)	(0.68)	
Cisco	0.001	3.014	-6.330	0.51%
	(2.37)	(0.30)	(-0.59)	
Disney	0.003	11.456	-16.849	1.06%
	(2.81)	(0.73)	(-1.21)	
Ebay	0.038	9.529	1.320	13.09%
	(7.40)	(1.14)	(0.23)	
GE	0.001	20.804	0.910	1.14%
	(3.58)	(2.61)	(0.07)	
IBM	0.019	24.185	-41.013	9.29%
	(6.96)	(2.19)	(-2.18)	
Intel	0.002	48.840	-4.830	5.60%
	(3.81)	(4.92)	(-0.47)	
Johnson & Johnson	0.020	13.573	-2.097	5.59%
	(5.39)	(1.35)	(-0.18)	
JP Morgan	0.001	11.732	5.424	0.71%
	(2.19)	(1.07)	(0.77)	
Merck & Co.	0.020	1.419	-9.697	1.91%
	(2.77)	(0.16)	(-0.81)	
Microsoft	0.007	1.860	-7.234	2.77%
	(4.46)	(0.15)	(-0.74)	
Pfizer	0.021	-1.296	-26.691	9.99%
	(5.63)	(-0.09)	(-1.89)	
UTC	0.044	8.872	-6.981	5.64%
	(5.83)	(0.90)	(-0.64)	
Verizon	0.001	-15.266	-4.399	0.52%
	(3.69)	(-0.97)	(-0.32)	
Wal Mart	0.001	0.425	-3.368	0.64%
	(2.44)	(0.03)	(-0.17)	
Yahoo	0.022	14.284	-17.784	14.49%
	(6.93)	(1.06)	(-3.73)	

Table 15: **Effect of Different News Categories on Probability of Daily Jump (Top 20 Firms) using All News from RavenPack Data**

This table reports coefficients from pooled logistic regression of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone by category for the top 20 news firms using RavenPack data for both novel and all news. The explanatory variables are the total number of news and absolute news tone reported on the RavenPack database each day. The sample period is from January 2000 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. The odds ratios are reported in brackets.

	$NewsCount_t$	$\lvert NewsTone_t \rvert$	$NewsCount_t$	$\lvert NewsTone_t \rvert$
	Top 20 Firms Total			
	Novel News		All News	
M&A	-0.022	0.032	0.042	-0.002
	(-0.50)	(0.99)	(1.62)	(-0.05)
	[0.978]	[1.032]	[1.043]	[0.998]
Analyst Ratings	0.254	0.061	0.218	0.077
	(8.82)	(1.90)	(8.50)	(2.41)
	[1.289]	[1.063]	[1.243]	[1.080]
Assets	0.085	-0.038	0.017	0.016
	(2.67)	(-0.77)	(0.44)	(0.39)
	[1.089]	[0.963]	[1.017]	[1.016]
Capital Structure	0.059	0.005	0.069	0.033
	(2.19)	(0.14)	(2.82)	(0.94)
	[1.061]	[1.005]	[1.071]	[1.034]
Credit Ratings	0.178	-0.126	0.122	-0.072
	(8.82)	(-2.91)	(5.10)	(-1.49)
	[1.195]	[0.881]	[1.130]	[0.931]
Earnings and Revenues	0.147	0.151	0.133	0.232
	(4.95)	(7.00)	(5.18)	(7.80)
	[1.158]	[1.163]	[1.142]	[1.261]
Marketing and Investor Relations	-0.112	0.056	-0.086	0.053
	(-1.93)	(1.63)	(-1.39)	(1.43)
	[0.894]	[1.058]	[0.917]	[1.055]
Labor Issues including Executive Turnovers	-0.012	0.010	0.036	-0.004
	(-0.30)	(0.39)	(1.13)	(-0.17)
	[0.988]	[1.010]	[1.037]	[0.996]
Products and Services	0.003	0.061	0.011	0.061
	(0.05)	(0.91)	(0.18)	(1.09)
	[1.003]	[1.063]	[1.011]	[1.063]
$R^2_{McFadden}$		7.65%		8.66%

Figure 1: **Raw & Filtered News Counts**

This figure plots the raw and filtered news counts from the first-stage Kalman filter for 4 representative firms. Raw news counts are plotted in blue and the filtered news counts are plotted in red for the in-sample period from January 2000 to December 2009.

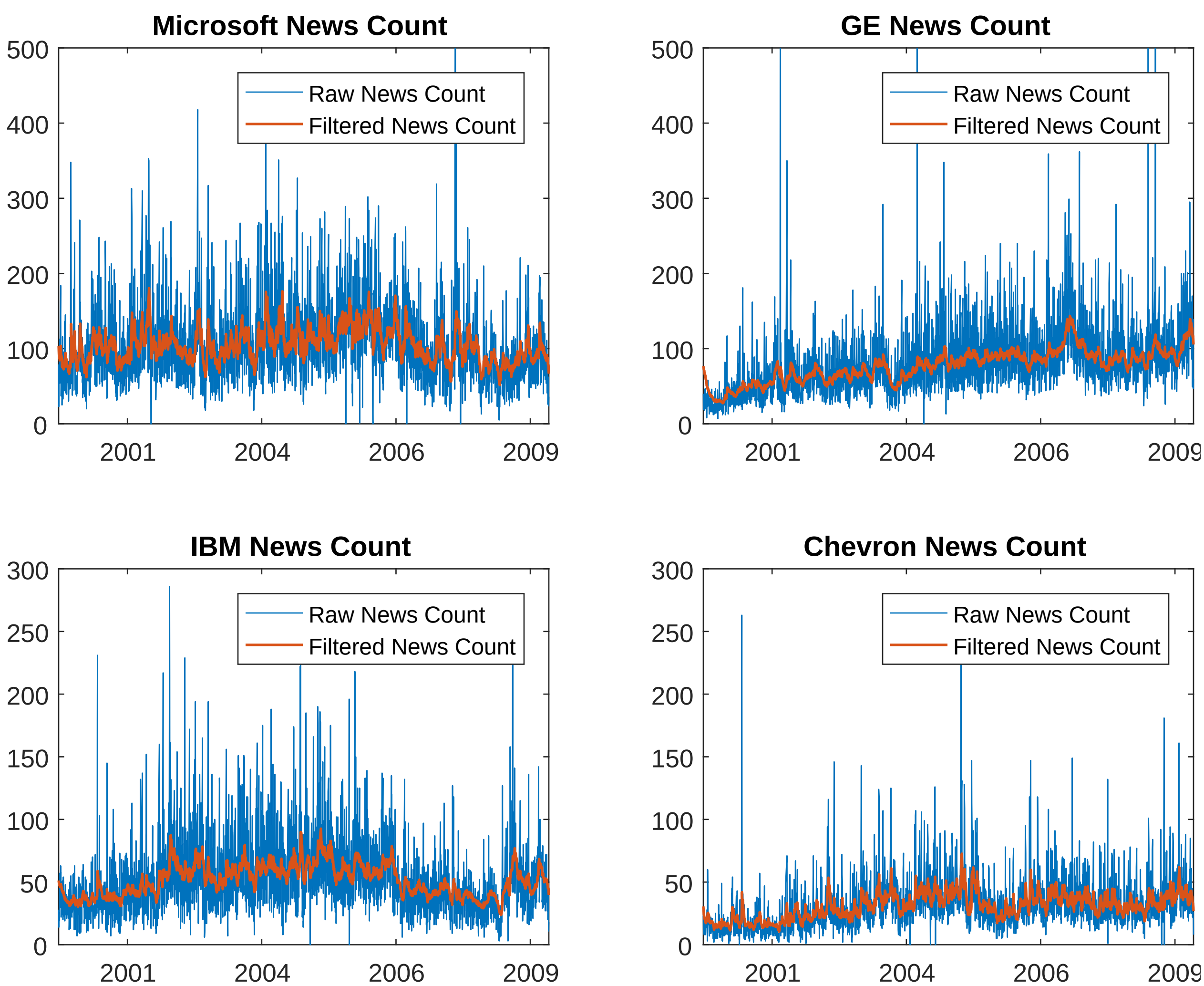

Table A.1: **Effect of News Counts on Probability of Daily Jump (Large Firms). 1980-2012**

This table reports coefficients from the pooled logistic regression of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for large firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panel B reports the odds ratios of each variable in brackets. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: Coefficient Estimates			
$NewsCount_t$	0.2390	0.2258	0.1839	0.1485
	(72.42)	(75.27)	(76.63)	(74.25))
$\|NewsTone_t\|$	0.0396	0.0317	0.0178	0.0123
	(13.20)	(12.19)	(9.89)	(8.79)
$UncWords_t$	0.0064	0.0066	0.0037	0.0025
	(2.06)	(2.54)	(2.18)	(1.92)
$\|Ret_{t-1}\|$	0.0504	0.0606	0.0729	0.0738
	(15.27)	(22.44)	(42.88)	(56.77)
N	9,426,014	9,426,014	9,426,014	9,426,014
$R^2_{McFadden}$	2.18%	1.76%	0.97%	0.60%
	Panel B: Odds Ratios			
$NewsCount_t$	[1.270]	[1.253]	[1.202]	[1.160]
$\|NewsTone_t\|$	[1.040]	[1.032]	[1.018]	[1.012]
$UncWrods_t$	[1.006]	[1.007]	[1.004]	[1.002]
$\|Ret_{t-1}\|$	[1.052]	[1.062]	[1.076]	[1.077]

Table A.2: **Effect of News Counts on Probability of Daily Jump (Medium Firms). 1980-2012**

This table reports coefficients from the pooled logistic regression of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for medium firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panel B reports the odds ratios of each variable in brackets. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: Coefficient Estimates			
$NewsCount_t$	0.1990	0.1886	0.1553	0.1283
	(62.19)	(65.03)	(67.52)	(64.15)
$\|NewsTone_t\|$	0.0118	0.0101	0.0099	0.0060
	(4.21)	(3.88)	(4.95)	(3.75)
$UncWords_t$	0.0132	0.0129	0.0077	0.0063
	(4.89)	(5.16)	(4.05)	(4.20)
$\|Ret_{t-1}\|$	0.0900	0.0994	0.1146	0.1250
	(29.03)	(38.23)	(60.32)	(78.13)
N	5,950,623	5,950,623	5,950,623	5,950,623
$R^2_{McFadden}$	1.36%	1.21%	0.85%	0.67%
	Panel B: Odds Ratios			
$NewsCount_t$	[1.220]	[1.207]	[1.168]	[1.137]
$\|NewsTone_t\|$	[1.012]	[1.010]	[1.010]	[1.006]
$UncWrods_t$	[1.013]	[1.013]	[1.008]	[1.006]
$\|Ret_{t-1}\|$	[1.094]	[1.104]	[1.121]	[1.133]

Table A.3: **Effect of News Counts on Probability of Daily Jump (Small Firms). 1980-2012**

This table reports coefficients from the pooled logistic regression of daily jump indicator defined using Lee and Mykland (2008) on daily news count and absolute news tone for small firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panel B reports the odds ratios of each variable in brackets. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: Coefficient Estimates			
$NewsCount_t$	0.1504	0.1463	0.1266	0.1074
	(50.13)	(52.25)	(57.55)	(56.53)
$\|NewsTone_t\|$	0.0102	0.0084	0.0067	0.0050
	(3.29)	(3.00)	(2.91)	(2.50)
$UncWords_t$	0.0113	0.0116	0.0072	0.0060
	(3.90)	(4.30)	(3.13)	(3.16)
$\|Ret_{t-1}\|$	0.0975	0.1077	0.1252	0.1359
	(33.62)	(43.08)	(69.56)	(84.94)
N	4,703,057	4,703,057	4,703,057	4,703,057
$R^2_{McFadden}$	0.79%	0.79%	0.69%	0.62%
	Panel B: Odds Ratios			
$NewsCount_t$	[1.162]	[1.158]	[1.135]	[1.113]
$\|NewsTone_t\|$	[1.010]	[1.008]	[1.007]	[1.005]
$UncWrods_t$	[1.011]	[1.012]	[1.007]	[1.006]
$\|Ret_{t-1}\|$	[1.102]	[1.114]	[1.133]	[1.146]

Table A.4: **Effect of News Counts on Daily Jump Size (Large Firms). 1980-2012**

This table reports coefficients from regressions of daily jump sizes conditional on the jump indicator being 1 on daily news count and news tone for large firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panels B and C report the results for positive jump sizes and negative jump sizes, respectively. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: All Jumps			
$NewsCount_t$	-1.38E-04	-1.38E-04	-8.52E-05	-1.10E-05
	(-6.17)	(-7.65)	(-8.95)	(-4.39)
$NewsTone_t$	0.8703	0.7094	0.4128	0.2723
	(39.85)	(42.62)	(48.97)	(53.81)
$UncWords_t$	0.4036	0.3087	0.1650	0.1081
	(14.66)	(14.70)	(15.47)	(16.85)
Ret_{t-1}	0.1610	0.1121	0.0062	-0.0603
	(19.02)	(17.37)	(1.84)	(-28.68)
N	117,683	172,332	445,746	894,227
R^2	1.88%	1.42%	0.63%	0.45%
	Panel B: Positive Jumps			
$NewsCount_t$	7.18E-04	5.57E-04	2.65E-04	1.52E-04
	(23.68)	(24.03)	(22.76)	(21.18)
$NewsTone_t$	0.2080	0.1765	0.1211	0.0953
	(9.12)	(10.41)	(14.37)	(18.79)
$UncWords_t$	0.3394	0.2703	0.1666	0.1147
	(11.92)	(12.78)	(15.76)	(17.96)
Ret_{t-1}	-0.0559	-0.0676	-0.0933	-0.1024
	(-6.99)	(-11.19)	(-29.02)	(-50.16)
N	70,755	103,729	261,417	507,169
R^2	1.05%	0.83%	0.62%	0.66%
	Panel C: Negative Jumps			
$NewsCount_t$	-3.74E-04	-3.61E-04	-2.09E-04	-1.48E-05
	(-24.40)	(-28.33)	(-30.92)	(-10.19)
$NewsTone_t$	0.4923	0.3953	0.2113	0.1311
	(26.09)	(27.16)	(29.05)	(30.47)
$UncWords_t$	0.0735	0.0231	-0.0339	-0.0327
	(3.09)	(1.25)	(-3.66)	(-5.97)
Ret_{t-1}	0.1481	0.1046	0.0474	0.0078
	(17.28)	(16.23)	(15.35)	(4.17)
N	46,928	68,603	184,329	387,058
R^2	4.44%	3.57%	1.66%	0.55%

Table A.5: **Effect of News Counts on Daily Jump Size (Medium Firms). 1980-2012**

This table reports coefficients from regressions of daily jump sizes conditional on the jump indicator being 1 on daily news count and news tone for medium firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panels B and C report the results for positive jump sizes and negative jump sizes, respectively. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: All Jumps			
$NewsCount_t$	-7.64E-05	-4.62E-05	1.34E-05	5.04E-05
	(-1.02)	(-0.70)	(0.30)	(1.52)
$NewsTone_t$	0.9727	0.8341	0.5463	0.3923
	(24.89)	(26.76)	(31.79)	(35.59)
$UncWords_t$	0.5848	0.4880	0.3105	0.2207
	(13.21)	(13.87)	(15.81)	(17.68)
Ret_{t-1}	-0.1291	-0.1434	-0.1947	-0.2392
	(-15.93)	(-22.40)	(-55.47)	(-103.51)
N	111,943	152,270	341,381	627,889
R^2	0.81%	0.82%	1.19%	1.87%
	Panel B: Positive Jumps			
$NewsCount_t$	4.59E-03	3.87E-03	2.59E-03	2.01E-03
	(28.02)	(29.47)	(34.70)	(39.61)
$NewsTone_t$	0.4846	0.3965	0.2789	0.2111
	(10.32)	(10.80)	(14.17)	(16.74)
$UncWords_t$	0.7644	0.6439	0.4602	0.3478
	(13.98)	(15.11)	(19.97)	(23.76)
Ret_{t-1}	-0.1127	-0.1064	-0.0966	-0.0984
	(-11.87)	(-14.19)	(-24.02)	(-36.70)
N	66,976	90,621	196,402	347,216
R^2	1.71%	1.46%	1.12%	1.00%
	Panel C: Negative Jumps			
$NewsCount_t$	-9.89E-04	-1.03E-03	-9.68E-04	-9.08E-04
	(-26.49)	(-30.11)	(-38.71)	(-46.04)
$NewsTone_t$	0.3787	0.3113	0.1708	0.1115
	(13.50)	(13.70)	(13.29)	(13.46)
$UncWords_t$	-0.1519	-0.1285	-0.1615	-0.1321
	(-4.93)	(-5.15)	(-11.28)	(-14.43)
Ret_{t-1}	-0.0348	-0.0453	-0.0794	-0.1048
	(-5.80)	(-9.72)	(-30.25)	(-60.90)
N	44,967	61,649	144,979	280,673
R^2	3.39%	2.99%	2.47%	2.53%

Table A.6: **Effect of News Counts on Daily Jump Size (Small Firms). 1980-2012**

This table reports coefficients from regressions of daily jump sizes conditional on the jump indicator being 1 on daily news count and news tone for small firms in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have the same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. Panels B and C report the results for positive jump sizes and negative jump sizes, respectively. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: All Jumps			
$NewsCount_t$	1.05E-03	9.95E-04	1.05E-03	1.09E-03
	(3.08)	(3.30)	(4.86)	(6.67)
$NewsTone_t$	1.6124	1.4163	0.8492	0.5662
	(25.18)	(26.68)	(27.36)	(27.60)
$UncWords_t$	1.1977	1.0504	0.5731	0.3589
	(18.07)	(19.15)	(17.52)	(16.52)
Ret_{t-1}	-0.2973	-0.3102	-0.3510	-0.3680
	(-42.27)	(-52.56)	(-95.91)	(-148.97)
N	121,748	155,887	309,766	532,965
R^2	1.97%	2.19%	3.12%	4.14%
	Panel B: Positive Jumps			
$NewsCount_t$	2.20E-02	1.65E-02	1.04E-02	9.01E-03
	(29.25)	(28.46)	(30.43)	(36.21)
$NewsTone_t$	0.4914	0.4840	0.3130	0.1929
	(5.80)	(7.09)	(7.99)	(7.46)
$UncWords_t$	0.6202	0.5994	0.4450	0.3141
	(7.05)	(8.47)	(10.68)	(11.27)
Ret_{t-1}	-0.2987	-0.2897	-0.2672	-0.2445
	(-34.08)	(-39.71)	(-58.88)	(-78.04)
N	68,339	88,120	171,735	286,060
R^2	2.98%	2.73%	2.55%	2.55%
	Panel C: Negative Jumps			
$NewsCount_t$	-4.53E-03	-4.74E-03	-4.97E-03	-4.53E-03
	(-29.72)	(-33.14)	(-42.77)	(-49.05)
$NewsTone_t$	0.4589	0.4022	0.2658	0.2059
	(11.42)	(11.78)	(13.03)	(14.96)
$UncWords_t$	0.0475	0.0405	-0.0366	-0.0228
	(1.15)	(1.15)	(-1.72)	(-1.59)
Ret_{t-1}	-0.0621	-0.0782	-0.1092	-0.1302
	(-13.17)	(-19.69)	(-44.21)	(-77.89)
N	53,409	67,767	138,031	246,905
R^2	3.17%	3.15%	3.32%	3.71%

Table A.7: **Effect of News Counts on Daily Jump Vol (by Size). 1980-2012**

This table reports coefficients from the regression of daily jump volatilities, defined as log of squared jump size conditional on the jump indicator being 1, on daily news count and news tone for three different size groups in the sample. The explanatory variables are the total number of news reported on the Factiva database each day and its news tone, standardized to have same mean and standard deviation across firms. The news tone measure is constructed first at each individual article level by counting the number of positive and negative words from Loughran and McDonald (2011), then they are aggregated by value-weighting scheme using total number of words in the article. The sample period is from January 1980 to July 2012. t-stats computed using standard errors clustered at individual firm levels are reported in the parentheses. All regression specifications include a constant term that is not reported for brevity.

	(1) $J99$	(2) $J95$	(3) $J_0 99$	(4) $J_0 95$
	Panel A: Large Firms			
$NewsCount_t$	0.0067	0.0061	0.0033	0.0002
	(26.79)	(28.60)	(24.46)	(5.84)
$\|NewsTone_t\|$	11.5030	9.6704	6.6358	5.0089
	(42.84)	(44.98)	(51.60)	(56.37)
$UncWords_t$	-4.4292	-3.5161	-2.2055	-1.7915
	(-13.87)	(-13.71)	(-14.27)	(-16.67)
$\|Ret_{t-1}\|$	13.6164	13.8720	15.2433	16.7690
	(123.86)	(156.04)	(272.20)	(408.29)
N	117,683	172,332	445,746	894,227
R^2	13.91%	14.19%	15.09%	16.11%
	Panel B: Medium Firms			
$NewsCount_t$	0.0150	0.0155	0.0150	0.0139
	(22.22)	(24.79)	(30.33)	(33.27)
$\|NewsTone_t\|$	11.8311	10.3098	7.7393	6.3025
	(31.36)	(32.79)	(38.02)	(42.32)
$UncWords_t$	-3.7901	-3.0885	-1.8134	-1.4306
	(-9.08)	(-8.91)	(-7.98)	(-8.67)
$\|Ret_{t-1}\|$	10.0322	9.8095	9.8196	10.8350
	(121.71)	(143.70)	(221.65)	(319.56)
N	111,943	152,270	341,381	627,889
R^2	13.46%	13.45%	13.55%	14.65%
	Panel C: Small Firms			
$NewsCount_t$	0.0696	0.0675	0.0638	0.0595
	(30.36)	(32.45)	(38.29)	(42.03)
$\|NewsTone_t\|$	8.3841	7.1931	4.7592	3.5004
	(18.16)	(18.44)	(18.66)	(18.43)
$UncWords_t$	-4.2548	-3.5735	-1.6269	-0.9285
	(-9.07)	(-9.03)	(-6.19)	(-4.73)
$\|Ret_{t-1}\|$	6.2920	6.3963	6.9289	7.1954
	(122.43)	(143.89)	(219.62)	(295.96)
N	121,748	155,887	309,766	532,965
R^2	12.24%	12.83%	14.23%	14.65%

Table A.8: **Parameter Estimates from the GJI-N Model (Diffusive Parameters)**
This table reports parameters associated with the diffusive dynamics of the GJI-N model estimated on daily returns and filtered news count. t-stats computed using the outer product of gradient method are reported in the parentheses. The sample period is from January 2000 to December 2009.

Company	μ	ω	α_1	β_1	$\alpha_{a,j}$	α_a
Amazon	1.30E-03	3.20E-13	-5.534	0.962	-2.875	2.988
	(2.08)	(0.00)	(-14.82)	(221.57)	(-4.14)	(7.76)
American Express	-3.88E-05	1.76E-18	-24.802	0.938	-0.713	22.913
	(-0.14)	(0.00)	(0.00)	(173.43)	(-3.31)	(0.00)
AT&T	6.72E-05	9.59E-07	-3.897	0.931	-1.885	1.891
	(0.24)	(2.24)	(-10.97)	(122.12)	(-3.27)	(5.20)
Bank of America	-1.21E-05	2.31E-07	-3.532	0.913	-1.206	1.846
	(-0.05)	(0.80)	(-13.80)	(129.96)	(-4.64)	(7.01)
Chevron	4.15E-04	5.85E-06	-4.499	0.917	1.807	2.167
	(1.47)	(5.17)	(-5.44)	(84.34)	(4.06)	(2.65)
Cisco	4.78E-04	2.64E-06	-5.630	0.940	-2.856	3.495
	(1.12)	(3.06)	(-4.65)	(129.28)	(-2.57)	(2.86)
Disney	2.40E-04	1.11E-06	-6.197	0.955	-1.647	3.708
	(0.67)	(2.91)	(-4.73)	(177.17)	(-3.35)	(2.79)
Ebay	8.60E-04	1.78E-06	-3.292	0.909	-2.600	1.417
	(3.13)	(3.26)	(-12.63)	(87.01)	(-3.05)	(5.28)
GE	-3.03E-04	8.64E-07	-5.298	0.933	-2.917	3.389
	(-1.03)	(2.66)	(-4.84)	(129.66)	(-3.46)	(3.13)
IBM	2.77E-06	6.74E-07	-5.032	0.927	-2.546	3.329
	(0.01)	(2.17)	(-5.95)	(109.62)	(-5.19)	(4.01)
Intel	-1.79E-04	1.87E-06	-14.078	0.954	-2.231	11.752
	(-0.41)	(2.69)	(-0.96)	(170.30)	(-3.29)	(0.80)
Johnson & Johnson	2.52E-05	7.60E-07	-4.418	0.895	-1.877	3.030
	(0.13)	(2.65)	(-7.23)	(99.36)	(-5.11)	(4.98)
JP Morgan	1.50E-04	9.79E-07	-3.968	0.921	-1.630	2.068
	(0.46)	(2.55)	(-11.20)	(121.82)	(-2.91)	(5.85)
Merck & Co.	1.82E-04	2.66E-06	-5.977	0.931	-2.490	3.919
	(0.54)	(3.95)	(-4.36)	(130.91)	(-7.74)	(2.88)
Microsoft	-5.52E-05	5.30E-07	-4.364	0.944	-2.410	2.136
	(-0.16)	(1.90)	(-12.63)	(163.97)	(-5.68)	(5.80)
Pfizer	-3.60E-04	1.04E-06	-4.395	0.919	-2.367	2.757
	(-1.25)	(2.24)	(-8.78)	(101.47)	(-4.73)	(5.48)
UTC	1.76E-04	1.80E-06	-5.138	0.936	-2.784	3.015
	(0.50)	(3.76)	(-5.30)	(138.05)	(-5.71)	(3.09)
Verizon	-8.44E-05	5.99E-07	-3.467	0.914	-1.516	1.650
	(-0.32)	(1.19)	(-12.92)	(98.93)	(-3.54)	(5.95)
Wal Mart	7.19E-05	8.42E-07	-3.942	0.915	-3.523	2.584
	(0.28)	(1.51)	(-10.61)	(86.70)	(-2.68)	(6.49)
Yahoo	4.90E-04	1.29E-06	-67.615	0.962	-3.228	65.155
	(0.89)	(1.76)	(0.00)	(235.52)	(-4.10)	(0.00)

Table A.9: **Parameter Estimates from the GARJI Model**

This table reports parameters associated with the benchmark GARJI model estimated on daily returns. t-stats computed using the outer product of gradient method are reported in the parentheses. The sample period is from January 2000 to December 2009.

Company	μ	ω	α_1	β_1	$\alpha_{a,j}$	α_a	γ_0	γ_1	γ_2	θ	δ	lgl
Amazon	3.42E-04	2.03E-06	-5.578	0.948	-3.050	3.458	0.042	0.128	0.196	-0.005	0.112	5,122.47
	(0.48)	(1.40)	(-12.50)	(143.22)	(-5.45)	(7.54)	(4.76)	(0.77)	(3.85)	(-0.42)	(9.65)	
Amex	-4.97E-04	1.13E-06	-5.117	0.939	-1.116	3.085	0.010	0.796	0.203	-0.008	0.027	6,370.70
	(-1.59)	(2.93)	(-4.72)	(128.47)	(-2.63)	(2.87)	(1.81)	(8.31)	(2.07)	(-1.67)	(5.48)	
AT&T	8.39E-05	1.04E-06	-4.018	0.935	-1.649	1.878	0.034	0.000	0.294	-0.004	0.028	6,758.82
	(0.29)	(2.28)	(-10.43)	(125.27)	(-2.58)	(4.75)	(3.16)	(0.00)	(1.65)	(-0.85)	(5.37)	
BoA	9.50E-04	1.89E-06	-3.107	0.897	-1.704	1.223	0.002	0.883	0.241	-0.007	0.058	6,664.36
	(3.54)	(4.06)	(-13.14)	(94.85)	(-4.11)	(5.36)	(2.64)	(20.53)	(2.99)	(-0.58)	(5.26)	
Chevron	2.10E-04	2.80E-06	-3.570	0.928	0.628	1.180	0.000	0.717	0.278	0.003	0.131	7,010.78
	(0.74)	(3.36)	(-10.15)	(105.37)	(1.22)	(3.38)	(0.73)	(5.10)	(0.43)	(0.04)	(0.20)	
Cisco	2.76E-05	2.91E-06	-5.855	0.940	-2.584	3.774	0.036	0.000	0.238	-0.003	0.052	5,853.41
	(0.06)	(3.09)	(-3.93)	(127.79)	(-2.78)	(2.52)	(3.08)	(0.00)	(2.82)	(-0.35)	(7.44)	
Disney	5.52E-04	1.07E-06	-5.170	0.947	-1.823	2.853	0.012	0.572	0.247	-0.003	0.060	6,496.25
	(1.59)	(2.13)	(-6.91)	(140.26)	(-3.78)	(3.76)	(2.42)	(3.67)	(2.31)	(-0.30)	(7.95)	
Ebay	6.46E-04	2.50E-06	-3.231	0.901	-2.428	1.412	0.024	0.152	0.410	-0.010	0.031	6,906.96
	(2.32)	(3.71)	(-12.43)	(77.78)	(-2.89)	(5.35)	(4.66)	(1.29)	(3.84)	(-1.67)	(5.11)	
GE	-9.60E-04	1.32E-06	-6.083	0.930	-2.329	4.306	0.032	0.262	0.115	0.000	0.027	6,791.45
	(-3.31)	(2.94)	(-2.42)	(117.46)	(-3.79)	(1.73)	(1.92)	(0.72)	(1.68)	(-0.07)	(8.06)	
IBM	1.79E-04	6.43E-07	-8.673	0.938	-1.900	6.717	0.014	0.773	0.123	-0.002	0.028	6,986.74
	(0.67)	(1.64)	(-0.30)	(113.74)	(-5.07)	(0.24)	(1.96)	(6.59)	(2.48)	(-0.65)	(9.04)	

Table A.9 Continued:

Company	μ	ω	α_1	β_1	$\alpha_{a,j}$	α_a	γ_0	γ_1	γ_2	θ	δ	lgl
Intel	-4.39E-05	1.54E-06	-7.681	0.956	-2.168	5.307	0.018	0.046	0.045	-0.008	0.068	5,867.12
	(-0.10)	(2.02)	(-0.58)	(144.60)	(-3.18)	(0.40)	(0.49)	(0.02)	(0.54)	(-0.48)	(5.68)	
J&J	1.38E-04	4.68E-07	-5.410	0.903	-0.997	3.877	0.099	0.372	0.183	-0.002	0.014	7,762.89
	(0.70)	(1.30)	(-3.82)	(94.85)	(-3.70)	(2.76)	(2.04)	(1.20)	(2.53)	(-1.80)	(8.48)	
JP Morgan	1.05E-04	1.44E-06	-4.036	0.929	-6.459	2.046	0.004	0.000	0.375	-0.003	0.113	6,254.27
	(0.34)	(3.59)	(-10.99)	(134.41)	(-1.55)	(5.56)	(1.93)	(0.00)	(2.98)	(-0.12)	(1.94)	
Merck & Co.	3.63E-05	2.33E-06	-5.557	0.934	-2.573	3.489	0.033	0.036	0.319	0.000	0.061	6,748.56
	(0.11)	(3.51)	(-6.15)	(143.44)	(-7.71)	(3.90)	(5.24)	(0.19)	(5.10)	(0.00)	(14.19)	
Microsoft	-5.16E-05	6.03E-07	-4.656	0.945	-2.371	2.440	0.004	0.895	0.045	-0.005	0.054	6,627.64
	(-0.15)	(1.82)	(-11.67)	(171.44)	(-5.68)	(5.84)	(1.12)	(9.28)	(1.53)	(-0.71)	(9.32)	
Pfizer	-2.72E-06	7.29E-07	-4.267	0.933	-2.664	2.284	0.014	0.506	0.321	-0.005	0.039	6,854.59
	(-0.01)	(1.39)	(-10.29)	(123.42)	(-5.33)	(5.28)	(3.88)	(4.60)	(3.50)	(-0.94)	(6.85)	
UTC	5.92E-04	1.91E-06	-4.342	0.930	-2.589	2.200	0.004	0.033	0.260	-0.004	0.088	6,834.48
	(1.98)	(3.28)	(-8.07)	(122.00)	(-6.60)	(4.03)	(1.54)	(0.05)	(0.77)	(-0.16)	(4.69)	
Verizon	2.35E-05	8.22E-07	-3.497	0.922	-1.228	1.454	0.033	0.614	0.489	-0.001	0.021	6,920.74
	(0.09)	(1.52)	(-11.80)	(92.75)	(-2.45)	(4.83)	(2.99)	(4.91)	(3.56)	(-0.55)	(7.07)	
Wal Mart	-4.65E-06	9.43E-09	-4.471	0.941	-1.344	2.454	0.129	0.304	0.554	0.002	0.014	7,027.04
	(-0.02)	(0.02)	(-8.03)	(122.41)	(-3.08)	(4.18)	(3.06)	(2.31)	(3.93)	(1.09)	(5.37)	
Yahoo	7.86E-06	2.05E-06	-5.719	0.951	-2.724	3.516	0.051	0.001	0.245	-0.001	0.081	5,250.38
	(0.01)	(1.81)	(-6.93)	(148.78)	(-5.04)	(4.29)	(5.01)	(0.01)	(3.81)	(-0.13)	(10.80)	