NBER WORKING PAPER SERIES

SO MANY JUMPS, SO FEW NEWS

Yacine Aït-Sahalia
Chen Xu Li
Chenxu Li

Working Paper 32746
http://www.nber.org/papers/w32746

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2024

The research of Chenxu Li was supported by the Guanghua School of Management, the Center for Statistical Sciences, the High-performance Computing Platform, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Ministry of Education) at Peking University, as well as the National Natural Science Foundation of China (Grant 72173003). The research of Chen Xu Li was supported by the National Natural Science Foundation of China (Grant 72203221). Chen Xu Li is also grateful for the financial support of the Bendheim Center for Finance at Princeton University and the School of Business at Renmin University of China. Thomson Reuters News Analytics provided access to its news data solely for academic purposes, without compensation or the right to review the research product. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

So Many Jumps, So Few News
Yacine Aït-Sahalia, Chen Xu Li, and Chenxu Li
NBER Working Paper No. 32746
July 2024
JEL No. G12,G14

ABSTRACT

This paper relates jumps in high frequency stock prices to firm-level, industry and macroeconomic news, in the form of machine-readable releases from Thomson Reuters News Analytics. We find that most relevant news, both idiosyncratic and systematic, lead quickly to price jumps, as market efficiency suggests they should. However, in the reverse direction, the vast majority of price jumps do not have identifiable public news that can explain them, in a departure from the ideal of a fair, orderly and efficient market. Microstructure-driven variables have only limited predictive power to help distinguish between jumps with and without news.

Yacine Aït-Sahalia
Department of Economics
Bendheim Center for Finance
Princeton University
Princeton, NJ 08540
and NBER
yacine@princeton.edu

Chen Xu Li
Renmin University of China
lichenxu@rmbs.ruc.edu.cn

Chenxu Li
Peking University
cxli@gsm.pku.edu.cn

A data appendix is available at http://www.nber.org/data-appendix/w32746

1. Introduction

The efficient markets hypothesis states that stock prices should fully reflect all available information at all times. When material new information becomes available, we expect prices to jump quickly to ensure that the information is incorporated into stock prices. Conversely, in a well functioning liquid market, prices should rarely jump in the absence of new information. This paper examines whether this is an adequate description of the reality on a granular level. We do so by studying the relationship between jumps in the high frequency prices of the Dow Jones Industrial Average (DJIA) stocks and firm-level, industry and macroeconomic news, in the form of low latency machine-readable releases from Thomson Reuters News Analytics (TRNA).

We begin by examining the relationship from news to price jumps. We find that relevant new information, both idiosyncratic and systematic, gets incorporated quickly into prices, as market efficiency suggests. This process can take several forms.[1] The news we are analyzing in this paper are by definition all public.[2] However, differences in speed among traders means that when news hit the wire in a machine-readable form, the fastest traders effectively possess private information over a short window of time, until slower traders can catch up and process the news at their own pace. The precise mechanism by which traders might trade to exploit their information, gradual or rapid, is not directly material to this paper. We are interested in the cumulative outcome, namely whether the price changes by a large amount (relative to its underlying volatility) over a short window of time. The windows of time we

[1]Sequential trade models provide a useful framework for analyzing how prices gradually incorporate new private information. In the Kyle (1985) model, an informed trader possesses private information about the future value of a stock, and optimally trades sequentially. Risk-neutral market makers try to infer from the aggregate order flow the information possessed by the insider. Because the order flow also includes uninformed noise traders, who trade solely for liquidity purposes, the informed trader has an opportunity to hide behind the noise traders and prices are not fully revealing. Instead, prices respond linearly to order flow and only gradually incorporate the information. In the Glosten and Milgrom (1985) model, informed traders have access to information about the new value of a stock, while uninformed traders do not. A specialist acts as a middleman between buyers and sellers. Traders arrive at the market sequentially and can observe the most recent transaction price. Informed traders are willing to buy or sell at a price that is closer to the new value of the asset than the current transaction price. Uninformed traders, on the other hand, are willing to buy or sell at the current transaction price. As informed traders enter the market and trade at prices that are closer to the new value of the asset, the price gradually moves towards it. In the model of Holden and Subrahmanyam (1992), multiple privately informed traders compete and trade aggressively, resulting in their private information being priced in rapidly.

[2]Since our focus in on public news and past price history only, the paper can speak only to the so-called "semi-strong" form of market efficiency.

consider, measured in seconds or minutes, are sufficiently long to allow for a sequential adjustment of prices from a series of transactions, resulting in a large cumulative price movement over a window of time, or for a single large price adjustment. So, from the perspective of semimartingale models of asset prices, we are not attempting to distinguish between a high drift burst concentrated in time (consistent with a rapid sequence of small price changes in the same direction, as in Christensen, Oomen and Renò (2022)) and a single price discontinuity. Over the short measurement window, we observe a large cumulative price change, which we label a "jump."

We find that the majority of news, whether idiosyncratic or systematic, lead to a price jump; that news that are scored by the language processing algorithm as more relevant for a given company, and/or scored as "newer" are more likely to lead to a jump in the time window immediately surrounding the news release; that news that are scored to have stronger sentiment for the firm are more likely to lead to a price jump; and that the price response to news happens quickly, in a matter of seconds. These results are broadly consistent with the hypothesis that the stock market is, to a large extent, informationally efficient.

However, when we analyze the reverse relationship, from price jumps to news, the situation is different. We separate stock price jumps into two categories: those that can be traced to new information about the firm itself, its industry or the aggregate economy, and those that cannot. Whereas we found that most relevant news lead to a jump, we now find that the vast majority of intraday jumps (at least 85%) do not have identifiable public news, whether idiosyncratic or systematic, that can explain them. Overnight jumps are quite different, with 96% of them being explained by news.

Why are so few intraday jumps explained by news? Our sample consists of the thirty DJIA stocks, which are among the most prominent firms in the U.S., and consequently very likely to have all their relevant news reported by TRNA and other news providers. It is possible that some of the jumps we observe are due to trading linked to news that are not publicly available, i.e., insider or other form of private information trading. (As we detail below, we design time windows to control for the possible advance leakage of soon-to-released public news.) We view this explanation as unlikely to rationalize the bulk of jumps without news: a firm is unlikely to generate significant information, private or public, at a pace consistent with the high rate of incidence of jumps we find in the data. Furthermore, as predicted by sequential trade models, an insider endowed with private information is expected to trade slowly to exploit and avoid revealing

the content of his or her information, making a rapid price adjustment, i.e., a jump, fairly unlikely, compared to the situation where fast traders acquire public information sooner than their slower counterparts and know that their window to act is short.

This leaves the architecture and process of trading as potential explanations for jumps without news. We explore various market microstructure considerations, such as the temporary disappearance of depth in the limit order book or other form of liquidity disruptions, the price impact of randomly arriving large noise orders, the reaction of market makers trying to protect themselves from adverse selection when transactions exhibit momentum, the strategic behavior by fast market makers able to anticipate the order flow, etc., as possible causes of these jumps without news. We then determine the set of features most predictive of jumps without news.

When comparing the distribution of jumps with news to that of jumps without news, we find that jumps with news, although substantially rarer, are more likely to have a larger price impact, less likely to be negatively autocorrelated, i.e., to be followed by subsequent price changes that undo some of their initial effect, and more likely to be clustered in time. But, importantly, even jumps without news do have some price impact that lasts minutes or more, so they are not immaterial.

This large proportion of jumps which do not have a clear cause in publicly available information, or in discernible large order arrivals, and the fact that they do not immediately self-correct, pose a challenge for the liquidity and stability of the stock market, both of which are, along with informational efficiency, key criteria for a well-functioning market.[3]

The paper is related to different strings of the literature, starting with classic studies investigating the informational efficiency of markets. Fama (1970) found few evidence to reject the efficient markets hypothesis in its strong, semi-strong, or weak forms. Roll (1988) found that most of the variation in the low-frequency returns of large stocks is difficult to explain using public firm-specific news. Roll (1984) found that orange juice futures prices imperfectly reflected changes in the relevant weather forecast for orange juice production. Cutler, Poterba and Summers (1989) found that major macroeconomic news had little explanatory power for the variance of aggregate stock

[3]A core mission of the Securities and Exchange Commission (SEC) is to "maintain fair, orderly, and efficient markets" (see https://www.sec.gov/about/mission). Although no precise definition of "orderly" is provided by the SEC, the term is widely understood to mean that securities regulations should seek to avoid "large, sudden price moves in individual stocks" that are not justified by fundamental information about the stock. Such an objective is also formally incorporated in the New York Stock Exchange (NYSE)'s requirements for its stock specialists.

prices.

Our finding that many stock price jumps cannot be explained by news on a scale of seconds or minutes is consistent with the established fact from the literature on excess volatility that prices are more volatile than is justified by subsequent movements in dividends on a time scale of many years (see, e.g., Shiller (1981), LeRoy and Porter (1981) and Lo and MacKinlay (1988)).

The paper is also related to the literature analyzing the impact of news on stock prices. Tetlock (2007) found that pessimistic media content induces downward pressure on market prices and higher trading volume; Tetlock, Saar-Tsechansky and Macskassy (2008) quantified the informational content in news using positive and negative words in news articles. Boudoukh et al. (2019) employed textual analysis and showed that the information contained in public news accounts for a large portion of idiosyncratic volatility. Dugast (2018) proposed a model to analyze the impact of news on volatility, trading volume, and order book imbalance.

Calomiris and Mamaysky (2019) examined 51 developed and emerging economies and found that the predictability of news is more significant for annual return and emerging markets. Jeon, McCurdy and Zhao (2022) estimated the impact of various news features on stock price jumps at the daily frequency and found that news play a role in explaining the occurrence of jumps and the distribution of jump sizes. On the other hand, Bajgrowicz, Scaillet and Treccani (2016) found that most news lead to a burst of volatility rather than jumps and Christensen, Oomen and Podolskij (2014) have argued that the price variation coming from jumps is overstated by standard tests. Chan (2003) found that stock returns exhibited momentum at the monthly frequency after public news, and that such momentum is particularly strong for bad news and illiquid stocks.

The literature has also examined how investors trade around news announcements. Hendershott, Livdan and Schürhoff (2015) compared institutional trading volume with news announcements and found that institutional traders appear capable of trading ahead of some news. On the other hand, Huang, Tan and Wermers (2020) found that institutional traders trade speedily on firm-specific news once they first become public and such news-motivated trading is profitable. Tetlock (2011) found that stale news can contain useless information and be aggressively traded by individual investors, explaining the return reversal the day after the release of stale news.

The paper is organized as follows. Section 2 describes how we combine the high

frequency price and news databases, and the model for stock prices we use to isolate jumps from the overall price path. Section 3 studies whether news lead to jumps. Section 4 examines the reverse direction, whether jumps can be traced back to news, and attempts to find determinants for the large proportion of jumps that are found to be without news. Section 5 concludes. Technical details and robustness checks are reported in the online supplement to the paper.

2. The Data, Model, and Jump Detection

2.1 The Data

We combine two data sources, high-frequency transaction prices and quotes from the NYSE's Trade and Quote (TAQ) database with machine-readable news from Thomson Reuters News Analytics (TRNA). Both prices and news are time-stamped at the millisecond level[4] and cover all Dow Jones Industrial Average (DJIA) stocks on all trading days from September 10, 2003 (when TAQ's coverage begins) to December 31, 2018. We continuously adjust the list of stocks to include the then-current list of DJIA constituents, resulting in more than thirty stocks in the sample due to additions and deletions to the index over the sample period. Table A.1 in the online supplement reports the list of stocks included in our sample, together with their sample periods, and any exclusions due to a short inclusion period in the index.

2.1.1 News Data

TRNA is a subscription-based news aggregation service that provides users with access to real-time news analytics and firm-specific sentiment data. TRNA scans news articles from Reuters and other news sources, including text data such as analysts' and brokers' recommendations, using a combination of natural language processing (NLP) and other machine learning algorithms. It then classifies their content, identifying which topics they cover, numerically scoring which companies they appear to be relevant for, whether the sentiment expressed is positive, negative, or neutral for that company, and the novelty of the news item. A separate database provides macroeconomic news in real time without numerical scoring. The news items are machine-readable, pushed to

[4]Post-2017 trades and quotes are nanosecond-stamped; we round them to milliseconds for consistency with the news releases.

subscribers in real time, using low latency servers located near the major financial centers around the globe. This service is often used by algorithmic traders, in conjunction with, or integrated into, their trading algorithms.[5]

Table 1 shows four examples of news records for Apple Inc., with the first two directly but the last two only indirectly relevant for Apple. Since a given same news item can be relevant for different companies, it is repeated in the database with different scores for different companies: except for the identification of the news item, i.e., the released time, headline, content and total word count, all the individual firm-specific scores such as relevance and sentiment are re-evaluated for each company.

Our analysis relies on the numerical assessment of news' relevance, sentiment, and novelty for given firm, as well as the word count used in the sentiment evaluation for a given company. We use these measures to construct progressively more stringent news filters, which we will discuss in detail in Section 3.1. First, the relevance of a news item to a specific firm is scored by a decimal number between 0 and 1. A news record with the highest relevance value to a company, i.e., 1, usually includes the name of this company in the headline. This is the case for the first and second examples in Table 1. The third (resp. fourth) example is directly related to Samsung (resp. Google), competitors of Apple in the smartphone market, and as a result indirectly relevant for Apple. TRNA assigns a relevance score of 0.10 (resp. 0.62) to Apple for the news in the third (resp. fourth) example to represent partial relevance.

Second, the sentiment of a news item is characterized by a three-dimensional vector, measuring the probabilities that the sentiment of the news item is positive, neutral, or negative for the given company, as assessed by natural language processing methods. The largest probability among the three entries determines the overall sentiment, which is summarized by a dummy variable taking value 1 for positive, 0 for neutral, and -1 for negative. For example, the first two news items in Table 1 are assessed as positive and negative respectively for Apple, with 84% positive and 82% negative probabilities respectively. The sentiment of the same news item can vary from one company to the next: positive firm-specific news for one company may be negative news for one of its competitors, or they could both be positive in the case of upbeat industry-specific news for two companies in the same industry.

Third, we use the number of words in the sentiment evaluation algorithm to capture

[5]von Beschwitz, Keim and Massa (2020) show that news analytics are used by traders for directional speculative bets based on the news' sentiment and that algorithmic trading combined with news analytics speeds up the response of the stock price and trading volume to the news.

the reliability of the sentiment probabilities: the more sentiment words in the news story, the more accurate the algorithm's sentiment evaluation is likely to be. Finally, the novelty of a news record is described by a five-dimensional vector. Each entry counts the number of similar news items over the previous 12 hours, 24 hours, 3 days, 5 days, and 7 days respectively. For example, a news item reporting a number relevant to a company may be repeated with further analysis or commentary, in which case we would expect the stock price to react mainly to the first occurrence of the news. In all four example records in Table 1, the novelty vector is $(0, 0, 0, 0, 0)$, indicating maximum novelty.

In addition to firm- and industry-specific news, we also take macroeconomic news into account. We include millisecond-stamped real-time news items concerning GDP growth, fiscal policy, labor market, trade, interest rate and monetary policy, housing market, and business conditions. Details are in Table A.2 of the online supplement. In total, we collect 35 types of macroeconomic news and have $6,107$ macroeconomic news items in the database. Unlike firm-specific news, macroeconomic news are not numerically scored by TRNA. However, since macroeconomic news are typically released at a much lower frequency, e.g., monthly, often on a pre-determined calendar, and their impact generally spreads across the entire market, we treat them as having relevance to all the companies in the DJIA, and only consider those with maximum novelty.

2.1.2 High Frequency Trade and Quote Data

The high-frequency trade and quote data are retrieved from the TAQ database. For each trading day, we collect all transaction data (timestamp, price, and volume) from 9:30 am to 4:00 pm, which we refer to as the "intraday" period. We collect the level-1 quote data (timestamp, price and sizes of the bid and/or ask) in a slightly longer time span for each trading day, from 9:00 am to 4:30 pm.

We pre-process the raw TAQ data using standard methods (see, e.g., Holden and Jacobsen (2014)): we match each transaction at the National Best Bid and Offer (NBBO) with the corresponding best quote(s) based on the timestamps and quote sizes. We then compute dollar-weighted and share-weighted effective spreads and quote midpoints, and sign the transactions using the buy-sell indicator of Lee and Ready (1991).

Prior to matching trades and quotes, we screen for data errors by eliminating crossed or locked quotes as well as transactions that have any irregular condition code in the

TAQ database, and further eliminate transactions that may otherwise be erroneous.[6] Once transactions and quotes are matched and signed, we apply a further filter that eliminates transactions with a price outside the NBBO range but a volume inside the quantity available at the best price (on the relevant side of the limit order book based on the sign of the transaction). From these millisecond-level transactions, we construct intraday log-returns at the baseline frequency of 5 seconds (which we will downsample as needed for robustness purposes) as well as overnight returns. Throughout the paper, "overnight" refers to the period from close to the following open over all market closures, including weekends and holidays.

2.2 Fitted and Abnormal Returns in a Multifactor Model

In order to quantify the impact of different types of news on a stock's price, we need to put some structure on stock prices: a model is needed before we can determine whether an individual stock's reaction to a given news item is as expected. To this aim, we employ a model with common factors and an idiosyncratic component, and classify the expected price impact of news as follows: (i) firm- and industry-specific news are reflected in the idiosyncratic component of the model; (ii) macroeconomic news are reflected via the common factors in the fitted component of the model; (iii) when both types of news are present together, both components of the model are potentially affected, i.e., the total return is impacted.

We model the dynamics of high-frequency prices in continuous-time, assuming that the log-price of stock Y_t follows the nonparametric factor model proposed in Aït-Sahalia,

[6]We only keep transactions whose Trade Correction Indicator is 1, Sales Condition is blank or @, and Trade Stop Stock Indicator is blank or N. Definitions of these variables are available from NYSE's website https://www.nyse.com/publicdocs/nyse/data/Daily_TAQ_Client_Spec_v4.0.pdf. We then exclude roundtrip jumps or bouncebacks formed by three successive transactions to avoid spuriously detecting of jumps later at a lower frequency. We say three successive transactions form a roundtrip jump or bounceback, if they satisfy the following two conditions: (1) the two log-returns are with opposite signs and both above thrice standard deviation (defined as the noise-robust realized volatility estimated over the previous trading day multiplied by the square root of time between two successive transactions, see Section 2.3 for more details), and (2) the absolute sum of two log-returns is less than 20% of the absolute value of the first log-return, suggesting an almost offsetting effect of two jumps caused by a single transaction, with subsequent transactions taking place at a price close to that prevailing before the isolated transaction. Although nothing in TAQ labels them as problematic, such transactions may have been recorded with an incorrect time or caused by other errors. In total, we eliminate 4.7% of the transactions across all stocks and dates. We choose the thresholds 3 and 20% as safeguards to eliminate a higher number of potentially incorrect transactions.

Jacod and Xiu (2023):

$$Y_t = Y_0 + \int_0^t \left(\beta_{s-}^C\right)^\intercal dX_s^C + \sum_{0\le s\le t} \left(\beta_{s-}^J\right)^\intercal \Delta X_s + Z_t, \tag{1a}$$

$$X_t = X_0 + \int_0^t b_s^X ds + \int_0^t \sigma_s^X dW_s^X + \int_0^t \int_E \delta^X(s,z)\mu^X(ds,dz), \tag{1b}$$

$$Z_t = Z_0 + \int_0^t b_s^Z ds + \int_0^t \sigma_s^Z dW_s^Z + \int_0^t \int_E \delta^Z(s,z)\mu^Z(ds,dz), \tag{1c}$$

where X is a $d-$dimensional multivariate covariate process representing common factors and Z represents the firm's idiosyncratic component.

This model is quite general. Both common factors and idiosyncratic components can have a continuous component (with stochastic volatility) and a jump component (with possibly infinite jump activity). X^C denotes the continuous component of the factor X, and ΔX_t denotes its jump, if any, at time t. The coefficient processes β^C and β^J represent the factor loadings on the continuous and jump components of X, respectively. Factor loadings β^C and β^J can be different. W^X (resp. W^Z) is a $d-$dimensional (resp. one-dimensional) standard Brownian motion. For the jump part, the Poisson random measure $\mu^X(ds,dz)$ (resp. $\mu^Z(ds,dz)$) is associated with the compensator measure ν^X (resp. ν^Z) of the form $\mu^X(ds,dz) = dt\otimes\lambda^X(dz)$ (resp. $\mu^Z(ds,dz) = dt\otimes\lambda^Z(dz)$) for some $\sigma-$finite measure λ^X (resp. λ^Z) on $\mathbb{R}^d$ (resp. $\mathbb{R}$). $\left(W^X,\mu^X\right)$ and $\left(W^Z,\mu^Z\right)$ are independent.

The model (1a)–(1c) is a continuous-time and high frequency generalization of discrete-time common factor models such as the Capital Asset Pricing Model (CAPM) and the models of Fama and French (1993) and Fama and French (2015). Its main feature compared to a discrete-time model is that it allows for a distinction between jumps and continuous components: in discrete-time, every price change is by definition discontinuous so just a distinction is not possible. The vector X of common factors consists of high frequency versions of the five Fama and French (2015) factors, i.e., the factors of market (MKT), value (HML), size (SMB), operating profitability (RMW), and investment (CMA). We also consider a continuous-time single factor CAPM and Fama-French three-factor model for robustness checks. (We find similar results.)

High frequency time series for the common factors are constructed using the method of Aït-Sahalia, Kalnina and Xiu (2020), applied to the sample of DJIA stocks. We then estimate the factor loadings for each stock using the continuous-time, high frequency,

version of the first pass Fama-MacBeth regression developed in Aït-Sahalia, Jacod and Xiu (2023). Continuous betas β^C are allowed to vary at the daily frequency and jump betas β^J at the annual frequency.[7] We use intraday returns in the estimation. For overnight stock and factor returns, we use the individual betas and the factor portfolios in effect on the last trading day before the overnight return.

Finally, based on the estimated factor model, we decompose the total return between t_{i-1} and t_i, i.e., $R_{[t_{i-1},t_i]} = Y_{t_i} - Y_{t_{i-1}}$ as

$$R_{[t_{i-1},t_i]} = R^{\text{fitted}}_{[t_{i-1},t_i]} + \epsilon_{[t_{i-1},t_i]}. \tag{2}$$

Here, $R^{\text{fitted}}_{[t_{i-1},t_i]}$ represents the fitted return from the factor model (1a)–(1c), i.e.,

$$\begin{aligned} R^{\text{fitted}}_{[t_{i-1},t_i]} &= \sum_{k=1}^{d} \hat{\beta}^C_{t,k} \left(X^{(k)}_{t_i} - X^{(k)}_{t_{i-1}} \right) 1_{\left\{ \left| X^{(k)}_{t_i} - X^{(k)}_{t_{i-1}} \right| \le u^{(k)}_t \right\}} \\ &\quad + \sum_{k=1}^{d} \hat{\beta}^J_{s,k} \left(X^{(k)}_{t_i} - X^{(k)}_{t_{i-1}} \right) 1_{\left\{ \left| X^{(k)}_{t_i} - X^{(k)}_{t_{i-1}} \right| > u^{(k)}_t \right\}}, \end{aligned} \tag{3}$$

where $\hat{\beta}^C_{t,k}$ (resp. $\hat{\beta}^J_{s,k}$) represents the daily (resp. annual) estimator of the kth entry of β^C_t on day t (resp. β^J_s in year s); $X^{(k)}_{t_i}$ represents the kth common factor; the $d-$dimensional vector $u_t = (u^{(1)}_t, u^{(2)}_t, \cdots, u^{(d)}_t)^\intercal$ contains the size thresholds used for differentiating continuous and jump components of X. $\epsilon_{[t_{i-1},t_i]}$ denotes the abnormal return of a stock as the change of Z_t, i.e., the residual return that cannot be linked to the common factors. Fitted and abnormal overnight returns from the prior closing price to the next opening price are computed based on the parameters in effect during the last trading day before the market closure. We provide more details in Section A.1 of the online supplement regarding factor construction, beta estimation, and thresholds selection.

As already noted, when firm- or industry-specific news (resp. macroeconomic news) are released, we expect the abnormal return $\epsilon_{[t_{i-1},t_i]}$ (resp. fitted return $R^{\text{fitted}}_{[t_{i-1},t_i]}$) of the stock to reflect the impact of the news. When both types of news are present together, we expect to observe the result in the total return $R_{[t_{i-1},t_i]}$, since both the common factors and idiosyncratic component are potentially affected.

[7] Fully general time-varying jump betas are not identifiable since only a single jump can happen at each instant. Multiple jumps are necessary to identify a jump beta.

2.3 Jump Detection at High Frequency

For each common factor, we only observe the total factor return X, not its continuous (X^C) and jump (ΔX) components separately, and similarly for each individual stock return. A decomposition is necessary to estimate the model (1a)–(1c) and compute abnormal returns in (2). The idea behind the estimation is the following: factor (or stock) returns that are larger than a multiple, say three or more, of their respective continuous volatility are unlikely to have been generated by the Brownian motion component of the model, and therefore classified as jumps. This is a standard method in high frequency econometrics, see, e.g., Lee and Mykland (2008) and Aït-Sahalia and Jacod (2014).

The first step is to estimate each factor (resp. stock) intraday volatility on a rolling basis day-by-day, using the standard noise-robust method of Zhang, Mykland and Aït-Sahalia (2005) that uses factor (resp. stock) intraday data within that day at 5s and combines volatility estimators at different frequencies ranging from 5s to 1mn. We then detect jumps, in the form of large realizations relative to what the Brownian component of the model is expected to generate given this volatility. We use noise-robust estimators RV_{t-1} of the time-varying volatility of a stock for each day $t-1$, and identify an intraday return $R_{[t_{i-1},t_i]}$ within day t as a jump if $\left|R_{[t_{i-1},t_i]}\right| > c \times RV_{t-1}\sqrt{\Delta}$, where $\Delta = t_i - t_{i-1}$ is the time interval between two successive observations, e.g., 5s in the baseline setting, and c is a fixed parameter specifying the multiple relative to standard deviation $RV_{t-1}\sqrt{\Delta}$ across the interval. Because the volatility estimator changes on a rolling daily basis with current market conditions, this method is robust to the possibility that periods with more news and/or jumps also experience higher volatility. For robustness purposes, we implement progressively more stringent truncation cutoffs starting from the baseline level $c = 3$ all the way to $c = 8$.

Overnight jumps are identified differently. Since the market is not trading continuously, there is no counterpart during the overnight hours to the continuous return volatility RV_t.[8] So instead of using a threshold relative to volatility as we did for detecting jumps during regular trading hours, we use an absolute threshold to detect overnight jumps. In the presence of a continuous component, the size of the returns due to the continuous component of the process are of order $\sqrt{\Delta}$ in probability. Increments

[8]Some limited after hours trading takes place on Electronic Communication Networks and other platforms but typically with much lower liquidity and trading volume than during the regular intraday hours on stock exchanges.

of any order greater than that are necessarily due to jumps. That includes a fortiori increments greater than a fixed finite threshold. We start with a threshold $c' = 2\%$ and for robustness purposes consider larger thresholds from 2.5% up to 4.5%.[9]

Before analyzing the data in detail, we begin with a brief description of the unconditional intraday distributions of news and detected jumps, respectively. For this purpose, we divide each regular trading session from 09:30 to 16:00 into 390 disjoint 1mn intervals. In the upper left panel of Figure 1, we count the number of news items in each intraday interval. In the lower left and lower middle panels, we report separately the total number of positive and negative jumps occurring in each time interval. First, it is clear that there are many more jumps than news during regular trading hours. Second, the total number of intraday jumps is typically larger at the beginning of the trading day, lower in the middle, and slightly larger towards the end of the trading day. Third, the proportions of positive and negative jumps are approximately the same. Fourth, we observe a periodic spike during the course of the day in the number of news items, resulting from firm-initiated news, which are slightly more likely to be released on the hour or half hour mark. This periodic behavior of news releases on the hour and half hour marks translates into a similar pattern for jumps, in line with the expectation that news often result in jumps.

Similar to the intraday plots, the upper right and lower right panels of Figure 1 count the number of overnight news items and jumps, respectively. As we will see below, even with a low $c' = 2\%$ threshold for detecting jumps, the vast majority of large overnight returns are associated with news. Increasing the threshold c' (i.e., considering only larger jumps) only makes this result stronger: larger overnight jumps are even more likely to be associated with news.

2.4 Time Windows Linking News and Jumps

In order to determine whether news results in a jump or not, and later on whether a jump is associated with news or not, we define windows of time surrounding the release of a news item, and the occurrence of a jump, respectively. Using the same contemporaneous window $[t_{i-1}, t_i)$ for both would not account for the possibility of

[9] An alternative would consist in using RV_{t-1} estimated during the intraday hours prior to the market closure and then detecting jumps on the basis of exceeding $c \times RV_{t-1}\sqrt{\Delta}$, where Δ is now the length of the overnight period. The counterfactual assumption implicit in this approach is that the market keeps trading continuously during the overnight period at the same rate per unit of time as it does during the day.

a slightly delayed response of the stock price to the news or conversely for potential news leakage and hence a price reaction slightly ahead of the public news release. So we allow for shifts of both the left and right boundaries of the time window: from news to jumps, we look for jumps both immediately before and after the release of the news; from jumps to news, we search for news both immediately before and after the occurrence of the jump. If the news-screening window of a jump is chosen as $[t_{i-1} - a\text{mn}, t_i + b\text{mn})$, the jump-screening window of a news item is accordingly chosen as $(t_{i-1} - b\text{mn}, t_i + a\text{mn}]$ so that news and jumps are symmetrically linked.

For robustness purposes, we consider different choices of time leads and lags by shifting the left and/or right boundaries of $[t_{i-1}, t_i)$ by up to 10mn; the baseline results shown in the paper correspond to a symmetrically shifted window of $[t_{i-1} - 2\text{mn}, t_i + 2\text{mn})$ (resp. $(t_{i-1} - 2\text{mn}, t_i + 2\text{mn}]$) for news (resp. jump) screening. For intraday news and jumps, if the left boundary is earlier than the market's opening time, i.e., 09:30:00, we shift the left boundary to the market's closing time on the previous trading day. If the right boundary is later than the market's closing time, i.e., 16:00:00, we further shift the right boundary to the market's opening time on the next trading day. For news and jumps happening during the overnight hours (including weekends and holidays), we shift the left boundary to the previous trading day's close time minus 2mn and the right boundary to the next trading day's open time plus 2mn.

3. From News to Jumps

3.1 News Filters

Only news items that are relevant to a company and contain new information are plausible candidates to move its stock price.[10] To this end, we consider five progressively more stringent news filters, described in the top panel of Table 2, for idiosyncratic firm and industry news. These filters are designed to isolate news that are more salient. The widest news filter, #1, consists of all news items that mention the specific company, irrespectively of the news item's relevance, sentiment or novelty.

In the remaining four filters, we impose some degree of novelty, relevance and sentiment. The degree of novelty allows us to exclude duplicated and/or stale news items, e.g., discussion, commentary or review of events that have already occurred. The

[10]For example, Tetlock (2011) found that, at the daily frequency, stock returns tend to respond less to stale news.

higher the degree of relevance, the more important the news item is to the company. The sentiment variables help in filtering news items with stronger, hence more reliable, sentiment. Compared with sentiment-neutral news, news with strong sentiment, either positive or negative, are in principle more likely to lead to less interpretative disagreement and consequently a clearer price reaction. The most stringent filter, #5, consists of news items that have the highest possible novelty, the highest possible relevance for the company in question and strong sentiment; specifically, the probability of either positive or negative sentiment is greater than 0.75, and the count of sentiment words in the news text is no less than 25.

As to the macroeconomic news, in the bottom panel of Table 2, we treat them as having relevance to all the companies in the sample, and select only those with maximum novelty: for instance, the first news item containing the release of a macroeconomic number. As to their relevance, we further divide the 35 types of macroeconomic news into two groups: the first contains all macroeconomic news, while the second is restricted to the (potentially most salient for asset valuation) news in the 4 categories of economic growth, unemployment, inflation and interest rates.

Table 3 reports a count of the number of firm-specific news items for each of the five news filters and each company. The table shows that the news dataset covers widely all the DJIA constituents, even under the strictest news filter #5. For a given choice of news filter, the number of idiosyncratic news items correlates with a company's size and the sample period: larger companies in the index tend to generate more information that is deemed newsworthy, attract more attention from the media, resulting in more news coverage, while naturally companies present in the sample for longer tend to also result in more news items.

We begin by documenting the impact of news on high frequency returns. First, we show that the presence of news in a time window leads to larger high frequency returns on average for all the individual firms in the sample, and in the aggregate, than the absence of news. The upper left panel in Figure 2 reports the results for all the individual companies in the sample in the form of the ratio of mean absolute returns with and without news; it is consistently above 1. In this panel, idiosyncratic (resp. macro) news correspond to firm-specific news filter #1 (resp. macro news filter #1). In the upper right panel of the figure, we show that as the news filters become tighter, the ratio of mean absolute returns increases. These patterns hold across the entire DJIA sample both in the aggregate (as shown in the upper right panel) and stock by stock

(not shown to save space).[11] Both firm-specific and macro news also lead to larger overnight returns in aggregate as shown in the lower right panel of Figure 2.

3.2 Impact of News on High Frequency Price Jumps

We now quantify the average number of jumps that follow a news item, then examine what features of the news item make it more likely that the stock price will jump in response, and by how much. Panel A of Table 4 shows that most intraday news lead to at least one jump, with an average of 1.26 (resp. 2.80) jumps (larger than 3 standard deviations) following the release of an idiosyncratic (resp. macro) news under filter #1; that number grows to 1.74 (resp. 7.18) jumps when the stricter news filter #5 (resp. #2) is used. The number of jumps following a news declines as progressively more stringent truncation cutoffs are used for jump detection.

It is natural to expect that not all news, even the most novel, relevant and clear in terms of sentiment, will lead to a jump. First, while we have measures of a news item's novelty, relevance and sentiment for a given company, we cannot capture its significance in terms of altering the fundamental valuation of this company. Some news may be significant on that dimension (such as a change of CEO, major litigation or regulatory announcements, etc.), but some may not be (such as an innocuous marketing press release), and their importance is likely context-specific. Second, we do not have a measure of how the news item compares to existing market expectations; some announcements, despite being novel, relevant, easy to interpret (strong sentiment) and even significant for valuation, may have already been anticipated and incorporated into the stock price, thereby not leading to a jump at the time of the news release.

Conversely, it is possible that a given news item leads to more than one jump in its screening window: the queuing mechanism by which orders get executed means that orders get processed sequentially, potentially resulting in successive price changes when traders with different speed levels react to the same news. We find in Panel A of Table 4 that the more relevant, novel and clear a news item is, the more likely we are to observe multiple jumps, with an expected number of jumps given news approaching two for idiosyncratic news filter #5 and exceeding seven for macro news filter #2. For overnight news, the difference between news items under different filters is starker as

[11] Robustness checks in Figure A.1 of the online supplement show similar patterns under alternative choices of return sampling frequency. Similar results also hold separately for positive and negative returns.

shown in Panel B of the table. Compared with 0.87 jumps for idiosyncratic news filter #1, idiosyncratic news filter #5 and macro news filter #1 (many of which are released during the overnight hours) leads to 5.14 and 3.50 jumps on average, respectively.

An interesting U-shape pattern in the average number of jumps detected also emerges in Table 4. Consider any of the columns for intraday news, corresponding to a fixed number of standard deviations. As we tighten the idiosyncratic news detection filter from #1 to #5, there are two counterbalancing effects at play. As the news become more salient, they are more likely to move the stock price, accounting for the increase in the average number of jumps recorded between filter #2 and filter #5. But between news filter #1 and news filter #2, the average number of jumps actually decreases. This is due to the fact that the news filter #1 is by design not very selective, and as a result some jumps detected as following a news item under filter #1 may not be related in a significant way to the actual news item.

We next quantify the proportion of news followed by jumps, i.e., $\mathbb{P}\left(\text{jump}|\text{news}\right)$, in Figure 3 under various settings on return sampling frequency, truncation cutoffs, jump-screening windows, and news filters. Generally, we find most overnight news induce jumps, while nearly half of intraday news induce jumps. This proportion declines, if one lowers the sampling frequency (emphasizing the importance of using high frequency data to relate news and jumps) and increases the magnitude of the jump.

To understand what drives these proportions, we consider several features of each news item, including the news' relevance, novelty, sentiment, and dummy indicators for macro news and overnight news. We then run the following Probit regression:

$$\Phi^{-1}\left(\mathbb{P}\left(\text{jump}|\text{news}_i\right)\right) = \alpha + \sum_j \beta_j \text{news feature}_{i,j} + \text{firm and year fixed effects} + \epsilon_i, \quad (4)$$

where Φ^{-1} is the inverse of the standard Normal cumulative distribution function and $\text{news feature}_{i,j}$ represents the jth news feature of the ith news item. The estimation results are reported in the top panel of Table 5. The estimators are all significantly positive in both univariate and multivariate regressions, and show that news that are more relevant to the company, have greater novelty and stronger sentiment, are important macro news, and are released in the overnight hours, all contribute to a higher likelihood of observing a price jump. Replacing the probability on the left hand side of (4) with the percentage jump size, and restricting the sample to intraday jumps, the middle panel shows that the same characteristics lead not only to a higher

likelihood of observing a jump but also to larger jumps on average. Restricting to overnight jumps results in less significant coefficients as shown in the bottom panel due to the reduction in sample size. Interestingly, the impact of macro news is less pronounced than that of firm-specific news during the overnight period.

Next, we investigate the delay (if any) in the jump response to news. In a liquid and efficient market, the response time of jumps to news ought to be short, so that the information contained in news is rapidly incorporated into the price. We focus on news that are linked to at least one jump within the jump-screening window $(t_{i-1} - 2\text{mn}, t_i + 2\text{mn}]$, where $[t_{i-1}, t_i)$ is a 5s interval including the news-releasing time t. Assume a jump arrives in the interval $[t_{j-1}, t_j) \subset [t_{i-1} - 2\text{mn}, t_i + 2\text{mn})$. We define the response time of a jump to the news as:

$$\text{response time} = \begin{cases} t_{j-1} - t, & \text{if } t_{j-1} \geq t_i \\ 0, & \text{if } t_{j-1} = t_{i-1} \\ t_j - t & \text{if } t_j \leq t_{i-1} \end{cases} . \tag{5}$$

If multiple jumps occur within the screening window $(t_{i-1} - 2\text{mn}, t_i + 2\text{mn}]$, we define the response time as the time gap between news and the jump closest to the news. By construction, the response time is between -2mn and 2mn.

Figure 4 shows the distribution of the response time of jumps to news year-by-year. The figure shows that the response time of jumps to news broadly shrinks over time, suggesting an improvement in informational efficiency over time. Moreover, the median response time declines to zero after 2008, meaning that the majority of jumps occur in the same 5s time interval as the news was released. While some jumps occur prior to the news, suggesting leakage or anticipatory ability, we consistently observe more jumps after the news (positive values in the figure) than prior to the news (negative values).

The conclusion from the analysis so far is that most news induce one (or more) price jump(s); that more relevant, more novel and more important news, as well as macro (systematic) and overnight news, make it more likely that a jump will take place; that the stronger the above news characteristics, the larger the jump; all of which happens quickly. So information regarding changes in the fundamentals affecting a company seem to be reflected efficiently in the stock price, both in terms of the magnitude of the effect and the speed with which it takes place.

But is this the full story? The question we now turn to is the reverse one, whether

all or most price jumps can be explained by news. And if not, what are the factors that can explain jumps that happen despite the absence of news?

4. From Jumps to News

In the reverse direction, we now separate detected jumps into two categories: those that have one (or more) news item(s) in the screening window that surrounds the jump vs. those that have none. To illustrate, Figure 5 shows two examples of jumps in Apple stock, one with and one without news. We then analyze the differences between these two categories of jumps. Of special interest is quantifying the rate of occurrence of jumps without news, and understanding the factors that make their occurrence more likely.

4.1 Jumps With and Without News

Section 3 showed that most news that are relevant and novel tend to lead to a jump. But the reverse is not true. We find that jumps with news are in fact a small subset of the aggregate set of jump realizations, except for jumps whose news-screening windows cover overnight periods, namely overnight jumps and jumps that happen at the very beginning and very end of a trading session.

Figure 6 plots the proportion of jumps that have news associated with them during the overnight period and across the 390 disjoint 1mn time intervals that make up the trading day, both stock-by-stock and in the aggregate. The figure shows that the proportion of jumps with news is generally below 15% throughout the trading day. Although subject to an increase around 10:00, 14:00, and in particular after 15:30, the proportion remains below 50% for most companies. The exceptions are the overnight period as well as the first and last two minutes of a regular trading session, where the proportions exceed 75% for most companies. Overnight jumps as well as intraday jumps before 09:32 or after 15:58 are associated with longer news-screening windows that include the overnight hours; by construction, we are more likely to find some news to attach to them.

Figure 7 reports the average proportion of jumps with news across stocks year by year. We observe an increase during the global financial crisis and a slight one overall, but the average remain fairly consistently below 20% over time. Figure 8 shows that the results are robust to changing the return sampling frequency, truncation cutoff

parameters, the length of the news-screening window, and the news filters; in each case, variation of the tuning parameters produces a monotonic pattern in the expected direction. The clear conclusion is that most intraday jumps are not news-driven, while most overnight jumps and/or intraday jumps at the very beginning and end of a trading session are.

While most jumps are not associated with news, jumps with news tend to be larger than jumps without news. Figure 9 compares histograms of the distribution of jump sizes (as a multiple of continuous volatility for intraday jumps, and in absolute size for overnight jumps) for jumps with and without news. The figure shows that most extremely large jumps, i.e., those in the far tails of the distribution, are news-driven.

4.2 Jumps Persistence

Next, we examine the persistence of the price impact of the two categories of jumps. In a perfectly efficient market, if a news item changes the market's valuation of a stock, we expect the effect of the price jump following the news to persist (at least until more news arrive). On the other hand, if the stock price jumped in the absence of news, we would expect the price to revert back to its previous level.

In the case of a jump following news, we measure the price impact of a jump as the cumulative return (CR) from a minutes prior to the jump to b minutes after the jump, defined as

$$\mathrm{CR}[-a,b] = \sum_{k\in[-a,b]} R_{[t_{i-1}+k,t_i+k]}, \tag{6}$$

where $R_{[t_{i-1}+k,t_i+k]}$ denotes the (log) return of the stock.

The results of this event study are in Figure 10. The left and right panels of the figure correspond to the impact of positive and negative intraday jumps, respectively.[12] Each curve in the figure plots the mean of $\mathrm{CR}[-10,b]$ as a function of b for the corresponding group of jumps. We find that jumps with news indeed have a more persistent price impact than jumps without news, but that even jumps without news have some persistent price impact that does not fully dissipate, at least up to 120mn after the jump.

[12]We do not consider overnight jumps, since the impact over time of a jump during the overnight vs. regular trading hours cannot be compared.

4.3 Jumps Clustering and Autocorrelation

Further, we find that jumps without news are more likely to occur in isolation. Conversely, jumps with news are more likely to be clustered over time, consistent with the explanation that traders are likely to react in quick succession to the same news (and to each other's trading), based on the speed with which they receive and process the news. Figure 11 shows the distribution of the jump counts, reporting the probability of $\mathbb{P}\left(k \text{ jumps}|\text{one jump in } [t_{i-1}, t_i)\right)$ as a function of the number of jumps k, when the jump in $[t_{i-1}, t_i)$ is without news, with firm- or industry-specific news, and with macro news, respectively. As shown in the figure, given a jump in $[t_{i-1}, t_i)$, we are more likely to observe additional jumps when the jumps are news-driven than when they are not.

Not only are jumps without news less likely to be clustered, but when they are, they are more likely to be negatively autocorrelated at the first lag. This means that, when a second jump follows a jump without news, the second jump is more likely to go in the opposite direction to that of the initial jump. We investigate in Table 6 the first order autocorrelation of consecutive jumps using the panel regression

$$\begin{aligned} \text{JSize}_{i+1} &= \beta_0 \text{JSize}_i + \beta_1 \text{JSize}_i \times 1_{\{\text{firm-specific news}\},i} + \beta_2 \text{JSize}_i \times 1_{\{\text{macro news}\},i} \\ &\quad + \text{firm and year fixed effects} + \epsilon_i, \end{aligned} \tag{7}$$

where JSize is the size of the jump. The indicator variable $1_{\{\text{firm-specific news}\}}$ (resp. $1_{\{\text{macro news}\}}$) takes 1 if the two consecutive jumps are associated with firm- or industry-specific news (resp. macro news), and the cross term JSize $\times 1_{\{\text{firm-specific news}\}}$ (resp. JSize$\times 1_{\{\text{macro news}\}}$) characterizes the difference between jumps without news and jumps with firm- or industry-specific news (resp. macro news). Table 6 shows the estimation results. First, the coefficient of JSize is negative, showing that consecutive jumps without news tend to be negatively autocorrelated. While this also holds for jumps with news due to the overall negative values of $\beta_0 + \beta_1$ (for jumps with firm-specific news) and $\beta_0 + \beta_2$ (for jumps with macro news), the coefficients of the two cross terms are positive, showing that compared to jumps without news, consecutive jumps with news are less negatively autocorrelated. So the impact of a jump with news is less counteracted by the jump afterward than a jump without news.

4.4 Trading Activity and Jumps

Next, we show in Figure 12 that jumps with news tend to be associated with higher levels of trading activity than jumps without news. This higher level of trading activity anticipates the jump for a longer period, and persists for longer as well, in the case of jumps with news. For jumps without news, we observe a smaller spike in trading volume immediately surrounding the time of the jump, and substantially less increase in trading activity outside of this short period: this concurs with the earlier evidence that jumps without news tend to be more isolated events than jumps with news.

To investigate this further, we compare in Figure 13 the trade-to-quote imbalance around jumps with and without news, which measures to what extent transactions consume the depth available in the limit order book. Since TAQ data does not provide quotes beyond the first level, an absolute value of imbalance is not meaningful. Instead, we measure the imbalance relative to one prevailing in time intervals without jumps. A relative value above 1 corresponds to a trade-to-quote ratio exceeding that prevailing in the absence of jumps, which is what we find surrounding jumps. Furthermore, this effect is substantially more pronounced and persistent for jumps with news than for jumps without news. We also note that jumps without news seem to occur following a period where the trade-to-quote imbalance is consistently high for a period of time. Then after the jump without news has occurred, the trade-to-quote imbalance quickly reverts to normal.

This evidence shows that the jumps without news that we identify are real events, in that they are associated with a short burst of higher trading volume, and a demand for liquidity that the supply of liquidity present in the limit order book at that point in time is unable to accommodate, with the price moving in response. But these jumps without news are not associated with abnormally high spikes in trading volume, which are lower than in the case of jumps with news, or with an abnormally high trade-to-quote imbalance.

4.5 Jumps Without News and Microstructure Features

What other factors could rationalize the large number of jumps that are not explained by new information that can affect the valuation of the stock? We saw in Figure 8 that jumps without news are substantially more prevalent in the intraday hours, when the market trades regularly, than they are during the overnight hours. This alone suggests

that explanations related to the trading process itself should be investigated.

We examine this question by running a Probit regression relating the probability that a jump is associated with news, $\mathbb{P}(\text{news}|\text{jump})$, to a number of market microstructure variables characterizing the stock's trading mechanism, as well as characteristics of the jump itself. These explanatory variables capture the stock return's dynamics, including volatility and momentum, its liquidity including the depth available in the limit order book and measures of price impact, the cost of trading including various measures of spreads, and any imbalance in the stock's transactions and/or quote availability. Table 7 describes the set of explanatory variables we use and their construction.

So symmetrical to the regression (4) for $\mathbb{P}(\text{jump}|\text{news})$, where we asked whether news led to a jump, we now ask whether jumps are associated with news in the form of the Probit regression:

$$\begin{aligned}\Phi^{-1}\left(\mathbb{P}\left(\text{news}|\text{jump}_i\right)\right) = \alpha + \beta x_i + \gamma_1 \text{AbsJSize}_i + \gamma_2 \text{JSign}_i \\ + \eta_1 1_{\{\text{overnight}\},i} + \eta_2 RV_{t-1} + \text{year fixed effects} + \epsilon_i. \end{aligned} \tag{8}$$

On the left hand side of (8), the probability $\mathbb{P}\left(\text{news}|\text{jump}_i\right)$ is the conditional probability that a return is driven by news, given that such return was classified as a jump. On the right hand side, x_i contains some or all of the explanatory variable introduced in Table 7 in the contemporaneous 5s window where the jump is detected.[13] We additionally control for jump size, jump sign, overnight period for news screening, realized volatility on the previous trading day, and year fixed effects in regression (8).

The results are in Figure 14. Each cell in the heatmap reports the estimator of β in regression (8) corresponding to a given explanatory variable and an individual stock. (Variables are standardized prior to estimation to make coefficients comparable across variables.) First, for the variables in the group of jump characteristics, the jump size explains the informativeness of a jump, while the jump sign generally does not: a jump with a larger jump size is more likely to be driven by news, consistent with our finding in Figure 9, irrespectively of its sign. Second, a jump with news is more likely accompanied by higher market liquidity. Although the average sizes on the bid and ask sides do not have a consistent impact, the total bid and ask sizes increase when the news is forthcoming. This is particularly so for news that are occurring

[13]Further inclusion of lagged explanatory variables to regression (8) does not change the estimation result substantially. If the explanatory variables x include the jump size or jump sign, the terms $\gamma_1 \text{AbsJSize}_i$ and $\gamma_2 \text{JSign}_i$ are removed from the regression.

on a pre-determined schedule, such as macroeconomic news releases. Third, a higher trading activity implies a higher likelihood of a jump with news, which is consistent with our finding in Figures 12 and 13 that the total trading volume and trade to quote imbalance around jumps with news are markedly higher. Fourth, the measures of trading cost, the two versions of realized spreads positively affect the likelihood of having a jump driven by news. Jumps with news have a stronger price impact, as shown in Figure 10. Finally, a comparison of the effect among the different groups suggests the limited explanatory power of quote data. Except for the total quote sizes, all the variables that are computed exclusively based on the quote data – BSize_avg, ASize_avg, AbsLOBImb, and QSpread – are shown to be insignificant or inconsistently impactful across the sample of individual stocks.

Interestingly, a higher transaction imbalance tends to a lower probability that the jump was associated with news. It may be that informed traders chop parent orders into numerous child orders and pool themselves with the uninformed ones so as to hide their information. Collin-Dufresne and Fos (2015) and Bloomfield, O'Hara and Saar (2015) found that informed traders make more use of hidden limit orders. In that case, the imbalance of active buys and sells may not be an indicator of underlying information.

In addition to regressions for individual stocks, we pool the jumps for various stocks together as follows

$$\begin{aligned}\Phi^{-1}\left(\mathbb{P}\left(\text{news}|\text{jump}_{i,k}\right)\right) = \alpha + \beta x_{i,k} + \gamma_1 \text{AbsJSize}_{i,k} + \gamma_2 \text{JSign}_{i,k} + \eta_1 1_{\{\text{overnight}\},i,k} \\ + \eta_2 RV_{t-1,k} + \text{firm and year fixed effects} + \epsilon_{i,k}. \qquad (9)\end{aligned}$$

The results are shown in Panel A of Table 8. As a robustness check, we also lower the sampling frequency from 5s all the way to 60s and find similar results, as shown in Table A.3 in the online supplement.

Finally, we examine whether microstructure features with explanatory power for $\mathbb{P}\left(\text{news}|\text{jump}\right)$ in sample also exhibit predictive power out of sample. For this, we examine the out-of-sample performance via the following Probit regression

$$\begin{aligned}\Phi^{-1}\left(\mathbb{P}\left(\text{news}|\text{jump}_{i,k}\right)\right) = \alpha + \beta x_{i,k} + \gamma_1 \text{AbsJSize}_{i,k} + \gamma_2 \text{JSign}_{i,k} \\ + \eta_1 1_{\{\text{overnight}\},i,k} + \eta_2 RV_{t-1,k} + \epsilon_{i,k}, \qquad (10)\end{aligned}$$

We treat the sample in year t as the in-sample data, and fit to it the panel regression

model (10). For computing the out-of-sample R^2, we additionally fit, to the same in-sample data, the following null panel regression model, which further excludes the explanatory variable $x_{i,k}$

$$\Phi^{-1}\left(\mathbb{P}\left(\text{news}|\text{jump}_{i,k}\right)\right) = \alpha + \gamma_1 \text{AbsJSize}_{i,k} + \gamma_2 \text{JSign}_{i,k} + \eta_1 1_{\{\text{overnight}\},i,k} + \eta_2 RV_{t-1,k} + \epsilon_{i,k}. \tag{11}$$

By construction, the only difference in model specification between the explanatory regression (10) and the null regression (11) lies in the inclusion of some variable x or not. Consequently, the improvement of the out-of-sample performance from the null model to an explanatory model exclusively suggests the predictability of the variable x.[14] We then compute the out-of-sample McFadden's pseudo R^2 of an explanatory model as

$$R^2_{\text{oos}} = 1 - \frac{\sum_{t=2003}^{2017} \ell_{\text{exp}}^{(t+1)}}{\sum_{t=2003}^{2017} \ell_{\text{null}}^{(t+1)}}, \tag{12}$$

where $\ell_{\text{exp}}^{(t+1)}$ (resp. $\ell_{\text{null}}^{(t+1)}$) represents the out-of-sample log-likelihood for the year $t+1$ under the explanatory (resp. null) model, respectively.

Panel B of Table 8 reports the out-of-sample R^2 in percentage for all the panel regressions. In a nutshell, the answer is yes to the question we raised at the beginning of the last paragraph: microstructure features with high (resp. low) explanatory power exhibit high (resp. low) predictive power out-of-sample. Indeed, the predictive power of individual explanatory variables can be categorized into three groups. Group 1 collects the variables with high predictive power, including AbsJSize, Breadth, Vol_all, and Turnover, each producing an out-of-sample R^2 of over 0.3%. Group 2 consists of variables with significant predictive power, including Immediacy, BSize_all, ASize_all, Vol_avg, Vol_max, Lambda, AbsTImb, TQImb, and two types of realized spreads, each of which induces an out-of-sample R^2 above 0.05% but below 0.3%. Group 3 incorporates variables with insignificantly predictive power, including the JSign, BSize_avg, ASize_avg, AbsLOBImb, Autocov, quoted spread, and two types of effective spreads,

[14] By regarding the sample in the next year, i.e., the year $t+1$, as the out-of-sample data, we compute the explanatory (resp. null) out-of-sample log-likelihood $\ell_{\text{exp}}^{(t+1)}$ (resp. $\ell_{\text{null}}^{(t+1)}$) based on the estimators obtained from the corresponding in-sample regression. Next, we roll one year forward and regard the sample of year $t+1$ (resp. $t+2$) as the in-sample (resp. out-of-sample) data. By repeating the previous steps, we compute two out-of-sample log-likelihoods $\ell_{\text{exp}}^{(t+2)}$ and $\ell_{\text{null}}^{(t+2)}$ for the year $t+2$. As rolling t forward year-by-year until the year 2017, we obtain a series of out-of-sample log-likelihoods $\ell_{\text{exp}}^{(t+1)}, \ell_{\text{exp}}^{(t+2)}, \ldots, \ell_{\text{exp}}^{(2018)}$ (resp. $\ell_{\text{null}}^{(t+1)}, \ell_{\text{null}}^{(t+2)}, \ldots, \ell_{\text{null}}^{(2018)}$) under the explanatory (resp. null) model.

each of which results in an out-of-sample R^2 below 0.05%.

Overall, we find that microstructure-driven variables have some limited predictive power to help distinguish between jumps with and without news. This predictive power is not large, and in particular jumps without news cannot be explained away as resulting simply from abnormally large transactions relative to the supply of liquidity that the market should not be expected to be able to absorb. This leaves us with the conclusion that we observe many "unexplained jumps" that cannot be justified by either the arrival of new information or sharp imbalances between the demand and supply for trading liquidity.

5. Conclusions

The paper related machine-readable news to high frequency jumps in both directions: from news to jumps and from jumps to news. News usually induce jumps in price: the probability $\mathbb{P}(\text{jump}|\text{news})$ is above 50% for both intraday and overnight news, even more so when the news is more relevant to the company, has greater novelty and stronger sentiment, is macro instead of firm-specific, and is released overnight.

In the reverse direction of jumps to news, jumps with and without news exhibit distinct market behavior. The distribution of jumps with news is notably sparse, with an average proportion of $\mathbb{P}(\text{news}|\text{jump})$ less than 20% among all jumps during the regular trading session. Jumps with news lead to stronger price impact, are accompanied by more trading activity, and are more clustered and less negatively autocorrelated. But even jumps that occurred in the absence of news have long-lasting price impact.

Market microstructure features have limited predictive power to explain the occurrence of jumps without news, distinguish them from jumps with news, or forecast their arrival. We intend to continue investigating the occurrence of jumps without news using different methods, including a reconstruction of the state of limit order book surrounding each one of these events to achieve a more granular understanding of each one of these jumps. For now, we have strong evidence that many jumps in the data are not an expected market reaction to news, leaving them as many examples of departures from the ideal of fair, orderly and efficient markets.

References

Aït-Sahalia, Yacine, and Jean Jacod. 2014. *High Frequency Financial Econometrics.* Princeton University Press.

Aït-Sahalia, Yacine, Ilze Kalnina, and Dacheng Xiu. 2020. "High Frequency Factor Models and Regressions." *Journal of Econometrics*, 216: 86–105.

Aït-Sahalia, Yacine, Jean Jacod, and Dacheng Xiu. 2023. "Continuous-time Fama-MacBeth regressions." Princeton University.

Bajgrowicz, Pierre, Olivier Scaillet, and Adrien Treccani. 2016. "Jumps in High-Frequency Data: Spurious Detections, Dynamics, and News." *Management Science*, 62: 2198–2217.

Bloomfield, Robert, Maureen O'Hara, and Gideon Saar. 2015. "Hidden liquidity: Some new light on dark trading." *The Journal of Finance*, 70: 2227–2274.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan, and Matthew Richardson. 2019. "Information, trading, and volatility: Evidence from firm-specific news." *The Review of Financial Studies*, 32: 992–1033.

Calomiris, Charles W, and Harry Mamaysky. 2019. "How news and its context drive risk and returns around the world." *Journal of Financial Economics*, 133: 299–336.

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

Christensen, Kim, Roel C.A. Oomen, and Mark Podolskij. 2014. "Fact or friction: Jumps at ultra high frequency." *Journal of Financial Economics*, 114: 576–599.

Christensen, Kim, Roel Oomen, and Roberto Renò. 2022. "The drift burst hypothesis." *Journal of Econometrics*, 227: 461–497.

Collin-Dufresne, Pierre, and Vyacheslav Fos. 2015. "Do prices reveal the presence of informed trading?" *The Journal of Finance*, 70: 1555–1582.

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

Dugast, Jérôme. 2018. "Unscheduled news and market dynamics." *The Journal of Finance*, 73: 2537–2586.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance*, 25: 34–105.

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

Fama, Eugene F., and Kenneth R. French. 2015. "A five-factor asset pricing model." *Journal of Financial Economics*, 116(1): 1–22.

Glosten, Lawrence R., and Paul R. Milgrom. 1985. "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders." *Journal of Financial Economics*, 14: 71–100.

Hendershott, Terrence, Dmitry Livdan, and Norman Schürhoff. 2015. "Are institutions informed about news?" *Journal of Financial Economics*, 117: 249–287.

Holden, Craig W., and Avanidhar Subrahmanyam. 1992. "Long-lived private information and imperfect competition." *The Journal of Finance*, 47: 247–270.

Holden, Craig W., and Stacey Jacobsen. 2014. "Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions." *The Journal of Finance*, 69: 1747–1785.

Huang, Alan Guoming, Hongping Tan, and Russ Wermers. 2020. "Institutional trading around corporate news: Evidence from textual analysis." *The Review of Financial Studies*, 33: 4627–4675.

Jeon, Yoontae, Thomas H McCurdy, and Xiaofei Zhao. 2022. "News as sources of jumps in stock returns: Evidence from 21 million news articles for 9000 companies." *Journal of Financial Economics*, 145: 1–17.

Kyle, Albert S. 1985. "Continuous auctions and insider trading." *Econometrica*, 53: 1315–1335.

Lee, Charles M.C., and Mark J. Ready. 1991. "Inferring trade direction from intraday data." *The Journal of Finance*, 46: 733–746.

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

LeRoy, Stephen F., and Robert D. Porter. 1981. "The present-value relation: Tests based on implied variance bounds." *Econometrica*, 49: 555–574.

Lo, Andrew W., and A. Craig MacKinlay. 1988. "Stock market prices do not follow random walks: Evidence from a simple specification test." *Review of Financial Studies*, 1: 41–66.

Roll, Richard. 1984. "Orange juice and weather." *The American Economic Review*, 74: 861–880.

Roll, Richard. 1988. "R^2." *The Journal of Finance*, 43: 541–566.

Shiller, Robert J. 1981. "Do stock prices move too much to be justified by subsequent changes in dividends?" *The American Economic Review*, 71: 421–436.

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

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

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy. 2008. "More Than Words: Quantifying Language to Measure Firms' Fundamentals." *The Journal of Finance*, 63: 1437–1467.

von Beschwitz, Bastian, Donald B Keim, and Massimo Massa. 2020. "First to "read" the news: News analytics and algorithmic trading." *The Review of Asset Pricing Studies*, 10: 122–178.

Zhang, Lan, Per A. Mykland, and Yacine Aït-Sahalia. 2005. "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data." *Journal of the American Statistical Association*, 100: 1394–1411.

Example 1
Timestamp: 2015-01-27 11:33:26.349
Headline: Apple to take first place in Chinese smart phone market for first time led by iPhone 6 and 6Plus units shipped - Research Firm Canalys
AssetID: 4295905573
Relevance: 1
SentimentClass: 1
(SentimentPositive, SentimentNeutral, SentimentNegative): $(0.8417540, 0.123264, 0.03498220)$
SentimentWordCount: 27
(NoveltyCount12Hours, 24Hours, 3Days, 5Days, 7Days): $(0, 0, 0, 0, 0)$
Example 2
Timestamp: 2015-06-03 10:01:37.982
Headline: US CPSC - Apple has received eight reports of incidents of the speakers overheating
AssetID: 4295905573
Relevance: 1
SentimentClass: -1
(SentimentPositive, SentimentNeutral, SentimentNegative): $(0.0556283, 0.125228, 0.81914400)$
SentimentWordCount: 15
(NoveltyCount12Hours, 24Hours, 3Days, 5Days, 7Days): $(0, 0, 0, 0, 0)$
Example 3
Timestamp: 2018-07-03 19:30:00.028
Headline: Samsung's Q2 profit seen flagging as smartphone innovation dries up
AssetID: 4295905573
Relevance: 0.1028690
SentimentClass: -1
(SentimentPositive, SentimentNeutral, SentimentNegative): $(0.2217280, 0.121574, 0.65669800)$
SentimentWordCount: 198
(NoveltyCount12Hours, 24Hours, 3Days, 5Days, 7Days): $(0, 0, 0, 0, 0)$
Example 4
Timestamp: 2015-06-06 12:05:28.540
Headline: No transaction fee for Google on mobile phone payments -WSJ
AssetID: 4295905573
Relevance: 0.6201740
SentimentClass: -1
(SentimentPositive, SentimentNeutral, SentimentNegative): $(0.1477200, 0.261691, 0.59058900)$
SentimentWordCount: 237
(NoveltyCount12Hours, 24Hours, 3Days, 5Days, 7Days): $(0, 0, 0, 0, 0)$

Table 1: Examples of TRNA news records for Apple Inc.

Note: AssetID is a unique identifier for the company, Apple, Inc., in this case. Relevance ranges from 0 to 1 and estimates how directly relevant the story is for the company. SentimentClass is a binary variable that classifies the news as either positive or negative for the company. The three-dimensional sentiment vector contains the respective probabilities that the news item's sentiment is positive, neutral and negative. SentimentWordCount reports the number of words matching that sentiment in the news story. The five-dimensional NoveltyCount vector measures the number of previously reported news items on the same topic over the past time intervals listed.

News filter	Description
Firm-specific news	
#1	Any news item that mentions the specific company.
#2	In addition to #1, news item having maximum novelty, i.e., the novelty vector is $(0, 0, 0, 0, 0)$.
#3	In addition to #2, news item having high relevance, i.e., relevance is greater than 0.5.
#4	In addition to #2, news item having maximum relevance, i.e., relevance is 1.
#5	In addition to #4, news item having strong sentiment, i.e., the probability of either positive or negative sentiment is greater than 0.75, and the sentiment words count in the text is no less than 25.
Macroeconomic news	
#1	A full set consisting of all 35 types of macroeconomic news. See Table A.2 of online supplement for details.
#2	A subset consisting of news on economic growth (GDP), inflation (CPI), unemployment rate, and interest rates (including FOMC meetings).

Table 2: News filters

Note: The five progressively more stringent firm-specific news filters are implemented separately for each firm. In firm-specific news filters #2–5, we exclude news whose headlines include the words "Buzz", "Reuters Insider", "Breaking Views", or "Wrap Up" as those consist mainly of summaries or commentaries on previously posted actual news. For macroeconomic news, only news with maximum novelty are included.

	Number of idiosyncratic news items				
Company	News filter #1	News filter #2	News filter #3	News filter #4	News filter #5
AA	19,004	9,394	4,504	4,126	511
AAPL	20,386	8,358	4,035	3,454	443
AIG	11,942	5,875	3,187	2,878	275
AXP	18,549	10,272	5,415	4,864	804
BA	72,933	36,543	20,047	16,833	2,281
BAC	50,411	26,359	11,973	9,630	884
C	44,528	21,343	8,672	7,239	846
CAT	22,397	10,416	4,878	4,433	495
CSCO	13,336	6,306	3,512	3,224	944
CVX	30,179	16,846	8,007	6,649	547
DD	17,189	9,527	5,402	5,008	752
DIS	31,478	19,202	8,926	7,439	1,038
GE	68,795	37,736	19,166	16,232	3,484
GM	49,524	25,067	14,215	11,611	1,211
GS	19,580	9,859	4,692	4,020	508
HD	17,668	9,479	4,212	3,911	314
HON	3,448	1,892	1,181	1,122	171
HPQ	27,626	14,335	6,209	5,680	594
IBM	36,068	20,997	11,893	11,019	3,715
INTC	37,482	19,001	9,314	8,307	1,280
JNJ	28,403	17,841	9,401	8,256	615
JPM	86,729	47,240	20,634	17,334	2,098
KFT	6,961	3,737	1,970	1,701	189
KO	21,965	12,772	6,164	5,257	809
MCD	19,443	10,166	4,859	4,486	457
MMM	13,912	7,528	4,695	4,205	891
MO	6,355	3,969	2,057	1,897	72
MRK	39,257	22,364	12,156	10,861	681
MSFT	63,699	33,842	15,631	13,684	2,916
NKE	5,588	2,587	859	786	88
PFE	41,796	22,235	11,116	9,939	783
PG	23,177	13,488	6,455	5,619	724
SBC	33,061	15,802	7,069	6,262	802
TRV	4,572	2,821	1,665	1,553	202
UNH	6,021	3,600	1,605	1,291	211
UTX	22,678	13,732	6,587	5,391	635
V	5,509	2,784	1,334	1,191	270
VZ	46,199	22,550	12,773	11,024	2,588
WMT	53,129	26,781	11,262	9,611	1,302
XOM	62,298	34,210	15,051	12,651	1,293

Table 3: Count of idiosyncratic news items by company

Note: Each cell in this table counts the total number of idiosyncratic news items over the whole sample period of the corresponding company. The definitions of the five news filters are given in the top panel of Table 2.

Panel A: Intraday news						
	⟵ smaller jumps				larger jumps ⟶	
Jump size cutoff	3	4	5	6	7	8
Idiosyncratic news: filter #1	1.26	0.56	0.27	0.15	0.092	0.059
Idiosyncratic news: filter #2	1.13	0.42	0.18	0.084	0.046	0.027
Idiosyncratic news: filter #3	1.28	0.53	0.25	0.13	0.079	0.055
Idiosyncratic news: filter #4	1.41	0.63	0.31	0.18	0.12	0.079
Idiosyncratic news: filter #5	1.74	0.77	0.41	0.25	0.16	0.11
Macro news: filter #1	2.80	1.30	0.64	0.34	0.19	0.12
Macro news: filter #2	7.18	4.79	3.29	2.33	1.68	1.22
Panel B: Overnight news						
	⟵ smaller jumps				larger jumps ⟶	
Jump size cutoff	2%	2.5%	3%	3.5%	4%	4.5%
Idiosyncratic news: filter #1	0.87	0.30	0.12	0.056	0.031	0.018
Idiosyncratic news: filter #2	2.33	1.16	0.57	0.31	0.18	0.11
Idiosyncratic news: filter #3	3.19	1.79	0.96	0.55	0.33	0.21
Idiosyncratic news: filter #4	2.88	1.85	1.22	0.82	0.57	0.40
Idiosyncratic news: filter #5	5.14	3.39	2.33	1.67	1.23	0.92
Macro news: filter #1	3.50	2.19	1.38	0.90	0.61	0.42
Macro news: filter #2	5.82	3.86	2.67	1.91	1.40	1.05

Table 4: Average number of jumps following news

Note: Panel A (resp. B) reports the average number of jumps led by one intraday (resp. overnight) news item, as a function of the category of news (macro or idiosyncratic under progressively more stringent news filters) and the jump size cutoff. In the row of news filter $\#i$, we only consider news included in filter $\#i$ but excluded from more stringent news filter, so that there is no overlapping news among different rows. The cutoff for intraday jumps is expressed as a multiple of the standard deviation c of the continuous variation of intraday returns, while for overnight jumps it is expressed as an absolute percentage value. To compute the number of jumps led by news, for each news item (intraday or overnight), we screen for the presence of jumps in the baseline jump-screening window $(t_{i-1} - 2\text{mn}, t_i + 2\text{mn}]$ surrounding the news in the window $[t_{i-1}, t_i)$. Intraday windows of time within two minutes of the opening and closing times include overnight returns, and the overnight windows symmetrically include the last two minutes of trading before close and the first two minutes after market open. In those cases, cutoffs for detecting jumps in the intraday and overnight periods respectively are adjusted in tandem from one column to the next. Within the screening window corresponding to the news item, each detected jump has a weight of 1. For each jump, if a single news item is present within the news-screening window of this jump, the weight of 1 is completely assigned to that news. When multiple news are present, we assign that jump's weight to the news with the highest filter, with macro news filter #1 (resp. #2) treated to be comparable with idiosyncratic news filter #3 (resp. #5). In the case of ties among multiple news items of the same filter, we treat the jump as being equally likely to have been led by each news item in the window and divide the jump's weight accordingly. Finally, for each news item, we sum the weight of each jump (a fraction of 1) detected in the jump-screening window as the number of jumps led by the news. We exclude any news item superseded by other news, if any, occurs: the number of jumps led by news is 0 if and only if there is no jump around the news. The number in each cell averages the results across corresponding news for all stocks and dates.

Dependent variable: $1_{\{\text{jump}\mid\text{news}\}}$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
News feature								
Relevance	0.752						0.395	0.396
	(181.8)						(87.1)	(87.4)
Novelty		0.771					0.180	0.183
		(83.5)					(18.8)	(19.1)
Strong sentiment			0.480				0.215	0.217
			(83.9)				(35.1)	(35.4)
$1_{\{\text{macro news}\}}$				0.893			0.532	
				(192.4)			(101.4)	
$1_{\{\text{macro news \#1}\}}$					0.809			0.498
					(165.9)			(91.5)
$1_{\{\text{macro news \#2}\}}$					1.537			0.845
					(106.3)			(54.6)
$1_{\{\text{overnight news}\}}$						1.229	1.076	1.066
						(300.9)	(251.6)	(247.8)
Number of observations	749, 107	749, 107	749, 107	749, 107	749, 107	749, 107	749, 107	749, 107
R^2	0.068	0.042	0.042	0.074	0.077	0.135	0.164	0.165
Dependent variable: Absolute intraday jump size (multiple of standard deviation)								
	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
News feature								
Relevance	1.219						0.248	0.252
	(80.6)						(16.3)	(16.5)
Novelty		2.173					0.131	0.140
		(63.6)					(3.9)	(4.2)
Strong sentiment			0.934				0.222	0.226
			(44.6)				(11.2)	(11.4)
$1_{\{\text{macro news}\}}$				0.752			0.216	
				(59.5)			(17.1)	
$1_{\{\text{macro news \#1}\}}$					0.549			0.177
					(40.5)			(13.1)
$1_{\{\text{macro news \#2}\}}$					1.653			0.395
					(64.8)			(16.2)
$1_{\{\text{overnight news}\}}$						2.692	2.597	2.584
						(266.8)	(241.5)	(237.7)
Number of observations	417, 655	417, 655	417, 655	417, 655	417, 655	417, 655	417, 655	417, 655
R^2	0.719	0.717	0.715	0.717	0.718	0.756	0.756	0.756
Dependent variable: Absolute overnight jump size (raw overnight return)								
	(17)	(18)	(19)	(20)	(21)	(22)	(23)	
News feature								
Relevance	0.0096					0.0116	0.0117	
	(7.2)					(8.5)	(8.5)	
Novelty		0.0094				0.0051	0.0054	
		(2.3)				(1.2)	(1.3)	
Strong sentiment			0.0021			0.0016	0.0016	
			(2.1)			(1.7)	(1.7)	
$1_{\{\text{macro news}\}}$				−0.0079		−0.0090		
				(−10.3)		(−11.6)		
$1_{\{\text{macro news \#1}\}}$					−0.0098		−0.0111	
					(−10.2)		(−11.5)	
$1_{\{\text{macro news \#2}\}}$					−0.0050		−0.0060	
					(−4.4)		(−5.2)	
Number of observations	10, 745	10, 745	10, 745	10, 745	10, 745	10, 745	10, 745	
R^2	0.725	0.723	0.723	0.726	0.726	0.728	0.729	

Table 5: Impact of news features on $\mathbb{P}\left(\text{jump}\mid\text{news}\right)$ and jump size

Note: The relevance score (between 0 and 1) is provided by TRNA for firm-specific news and set at 1 for macro news. For firm-specific news, we first winsorize NoveltyCount7Days at its 99% quantile and then scale it on $[0, 1]$. The variable Novelty in the regression is then defined as 1 minus the scaled NoveltyCount7Days, so that greatest novelty equals 1. Macro news are already selected exclusively as having maximum novelty equal to 1. For firm-specific news, strong sentiment is defined as the absolute difference between TRNA's SentimentPositive and SentimentNegative. For macro news, strong sentiment is set at 0.5. The indicator variables $1_{\{\text{macro news}\}}$, $1_{\{\text{macro news \#1}\}}$, $1_{\{\text{macro news \#2}\}}$, and $1_{\{\text{overnight news}\}}$ take values 1 for any macro news, macro news exclusively from filter #1 but not filter #2, macro news from filter #2, and any overnight news, respectively, and 0 otherwise. Regressions (1)–(8) concern the indicator variable $1_{\{\text{jump}\mid\text{news}\}}$ (which returns 1 if the number of jumps led by the news computed according to the note of Table 4 is greater than zero) as the dependent variable, corresponding to the Probit regression (4). We exclude superseded news items from all regressions. Regressions (9)–(16) (resp. (17)–(23)) concern the jump size; they focus on a subsample of news, where for each news item the number of jumps led by the news item is greater than zero, i.e., $1_{\{\text{jump}\mid\text{news}\}} = 1$, and the closest jump that has a nonzero weight to the news item is an intraday (resp. overnight) jump. The dependent variable turns to be the absolute jump size of the closest jump, measured as absolute multiple of standard deviation of the continuous variation (resp. raw overnight return) of intraday (resp. overnight) returns and winsorized at its 99% quantile. Regressions (9)–(23) are regular linear regressions. Numbers in parentheses report z-statistics (resp. t-statistics) for the Probit (resp. regular linear) regressions.

Measure of jump size	Jump return	Jump size multiplier
JSize	-0.244 (-598.0)	-0.158 (-387.1)
JSize$\times 1_{\{\text{firm-specific news}\}}$	0.090 (9.5)	0.023 (22.1)
JSize$\times 1_{\{\text{macro news}\}}$	0.101 (67.4)	0.033 (23.5)
Number of observations	$7,102,441$	$7,102,441$
R^2	0.059	0.024

Table 6: Autocorrelation of consecutive jumps

Note: For each cluster of jumps identified in Figure 11, we run the panel regression (7) for pairs of consecutive jumps with numbers in parentheses reporting the t-statistics. In the left column, JSize is measured as the total log-return, while in the right column JSize is expressed as a multiple of the continuous standard deviation. Both JSize_i and JSize_{i+1} are winsorized at 1% and 99% percentiles before running the regressions.

Variable	Abbreviation	Computational details	Data used
Jump characteristics			
Absolute jump size	AbsJSize	Absolute multiple of the standard deviation for intraday jumps.	T
Jump sign	JSign	An indicator variable taking 1 for positive jumps.	T
Liquidity			
Immediacy	Immediacy	Average time between successive transactions in $[a, b)$.	T
Breadth	Breadth	Number of transactions in $[a, b)$.	T
Average best bid sizes	BSize_avg	Average of all best bid sizes in $[a, b)$.	Q
Average best ask sizes	ASize_avg	Average of all best ask sizes in $[a, b)$.	Q
Total best bid sizes	BSize_all	Total best bid sizes in $[a, b)$.	Q
Total best ask sizes	ASize_all	Total best ask sizes in $[a, b)$.	Q
Absolute Limit order book imbalance	AbsLOBImb	Absolute value of (BSize_all−ASize_all)/(BSize_all+ASize_all).	Q
Transaction impact			
Total volume	Vol_all	Total volume traded in $[a, b)$.	T
Average volume	Vol_avg	Average volume traded in $[a, b)$.	T
Maximum volume	Vol_max	Maximum volume traded in $[a, b)$.	T
Turnover rate	Turnover	Sum of volume times price for all transactions in $[a, b)$, divided by the market cap of the stock, where the market cap is the average transaction prices in $[a, b)$ times the number of shares outstanding on the last trading day.	T
Kyle's lambda	Lambda	Absolute difference between prices of the first and last transactions in $[a, b)$ divided by Vol_all.	T
Autocovariance	Autocov	First order autocovariance of log-returns in $[a, b)$, with log-returns being the difference of two successive log-prices.	T
Absolute transaction imbalance	AbsTImb	Absolute difference between total buy size and total sell size divided by Vol_all, where the type of each transaction, i.e., buy or sell, is determined by following Lee and Ready (1991).	T
Trade to quote imbalance	TQImb	Total buy (resp. sell) size divided by ASize_all (resp. BSize_all) in $[a, b)$ for positive (resp. negative) jumps.	T & Q
Cost			
Time-weighted average of percentage quoted spreads	QSpread	Time-weighted average of all percentage quoted spreads in $[a, b)$, where the percentage of quoted spread is the difference between the best ask price and best bid price divided by the midpoint, i.e., the average of the best bid and best ask prices.	Q
Dollar-weighted average of percentage effective spreads	ESpread_D	Dollar-weighted average of all percentage effective spreads in $[a, b)$. Here, the kth percentage effective spread is $2D_k(P_k - M_k)/M_k$, where D_k is the Lee and Ready (1991) buy-sell indicator; P_k and M_k are the kth transaction price and midpoint, respectively.	T & Q
Share-weighted average of percentage spreads	ESpread_S	Share-weighted average of all percentage effective spreads in $[a, b)$, where the share of the kth effective spread is the volume of the kth trade.	T & Q
Absolute dollar-weighted average of percentage realized spreads	AbsRSpread_D	Absolute value of dollar-weighted average of all percentage realized spreads in $[a, b)$. Here, the kth percentage realized spread is $2D_k(P_k - M_{k+5})/M_k$, where D_k, P_k, and M_k are the same as those for computing ESpread_D, and M_{k+5} is the midpoint 5mn after the midpoint M_k.	T & Q
Absolute share-weighted average of percentage realized spreads	AbsRSpread_S	Absolute value of share-weighted average of all percentage realized spreads in $[a, b)$, with the share representing the volume of the corresponding trade.	T & Q
Control variable			
Coverage of overnight period	$1_{\{\text{overnight}\}}$	An indicator variable taking value 1 if the news-screening window of a jump includes an overnight period.	T
Realized volatility	RV	Realized volatility on the last trading day.	T

Table 7: Explanatory and control variables for the probability $\mathbb{P}(\text{news}|\text{jump})$ that a jump is associated with news

Note: $[a, b)$ represents any arbitrary intraday time interval. In $[a, b)$, if the number of transactions is less than 2 (resp. 3), we treat Lambda as missing and Immediacy as $b - a$ (resp. Autocov as missing). T and Q in the last column represent trade and quote data, respectively. To exclude extreme observations, we winsorize all the continuous explanatory variables: for the unbounded variable Autocov, we winsorize at the 1% and 99% percentiles; for the variables that are left bounded but not right bounded, such as Breadth and Turnover, we only winsorize at the 99% percentile to exclude extremely large observations; for bounded variables Immediacy, AbsLOBImb, and AbsTImb, we apply no winsorization. We fill in all missing values with the median of the data for the same stock in the same year. We centralize and standardize the data of each continuous explanatory variable so that all these variables have zero median and unit variance, thereby making the estimators of different variables directly comparable.

Panel A: In-sample estimation result							
Variable	Estimator (z-stat)	Variable	Estimator (z-stat)	Variable	Estimator (z-stat)	Variable	Estimator (z-stat)
Jump char.		Liquidity		Transaction impact		Cost	
AbsJSize	0.0816	Immediacy	−0.0625	Vol_all	0.0846	QSpread	0.0004
	(155.6)		(−107.6)		(166.7)		(0.70)
JSign	0.0034	Breadth	0.1023	Vol_avg	0.0284	ESpread_D	0.0138
	(3.16)		(199.0)		(53.9)		(22.6)
		BSize_avg	−0.0012	Vol_max	0.0486	ESpread_S	0.0138
			(−2.17)		(95.0)		(22.6)
		ASize_avg	0.0024	Turnover	0.0847	AbsRSpread_D	0.0607
			(4.43)		(166.9)		(109.2)
		BSize_all	0.0408	Lambda	−0.0452	AbsRSpread_S	0.0607
			(77.8)		(−70.7)		(109.2)
		ASize_all	0.0433	Autocov	0.0093		
			(82.6)		(14.5)		
		AbsLOBImb	0.0036	AbsTImb	−0.0490		
			(6.5)		(−84.6)		
				TQImb	0.0513		
					(97.5)		
Panel B: Out-of-sample performance							
Variable	R^2_{oos}	Variable	R^2_{oos}	Variable	R^2_{oos}	Variable	R^2_{oos}
Jump char.		Liquidity		Transaction impact		Cost	
AbsJSize	0.39	Immediacy	0.15	Vol_all	0.44	QSpread	0.022
JSign	−0.0001	Breadth	0.62	Vol_avg	0.056	ESpread_D	0.0087
		BSize_avg	0.026	Vol_max	0.15	ESpread_S	0.0087
		ASize_avg	0.030	Turnover	0.44	AbsRSpread_D	0.16
		BSize_all	0.12	Lambda	0.072	AbsRSpread_S	0.16
		ASize_all	0.15	Autocov	0.0050		
		AbsLOBImb	0.0008	AbsTImb	0.10		
				TQImb	0.19		

Table 8: Explanatory and predictive power of market microstructure variables for $\mathbb{P}(\text{news}|\text{jump})$

Note: Panel A shows the in-sample estimation results of regressions (9). Numbers in parentheses report the corresponding z-statistics. Panel B shows the predictive power of market microstructure variables in terms of out-of-sample R^2_{oos} in percentage defined in (12).

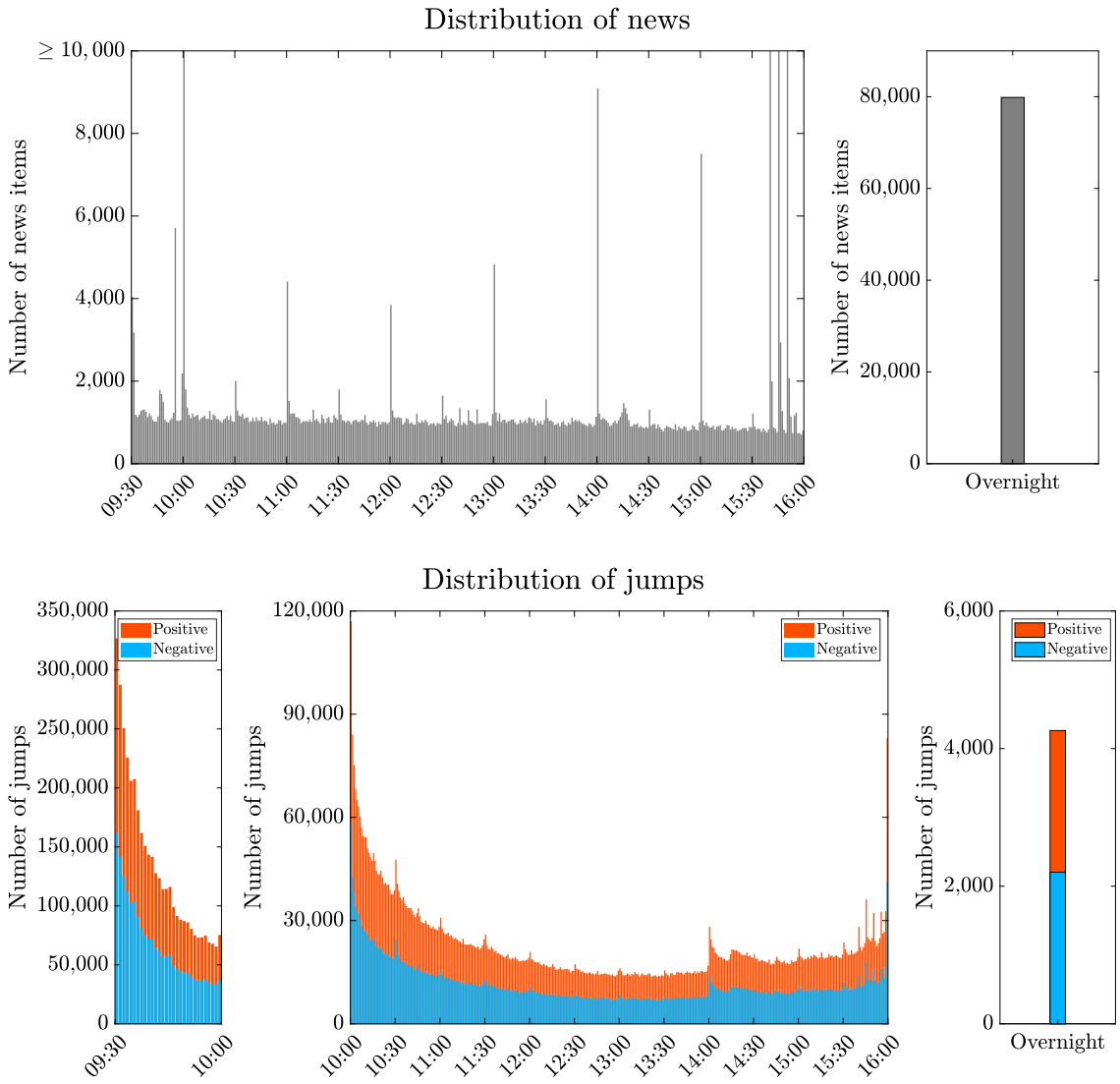

Figure 1: Distributions of intraday and overnight news and jumps

Note: The top two (resp. bottom three) panels show the distribution of news and jumps. In the upper left (resp. right) panel, each vertical bar counts the number of intraday news items in a corresponding 1mn time interval between 09:30 and 16:00 (resp. during the overnight period from market close to open) across all stocks. In the three panels at the bottom, each red (resp. blue) bar counts the number of positive (resp. negative) jumps in the corresponding time interval. To show the pattern clearly, We split the intraday period into two parts (09:30-10:00 and 10:00-16:00). The proportions of intraday 5s returns identified as jumps are respectively 2.3%, 0.90%, 0.41%, 0.22%, 0.13%, and 0.080%, as the cutoff c takes integer values 3, ..., 8. The proportion of overnight returns identified as jumps are respectively 3.7%, 2.4%, 1.6%, 1.2%, 0.93%, and 0.74% as the cutoff c' takes values 2%, ..., 4.5%.

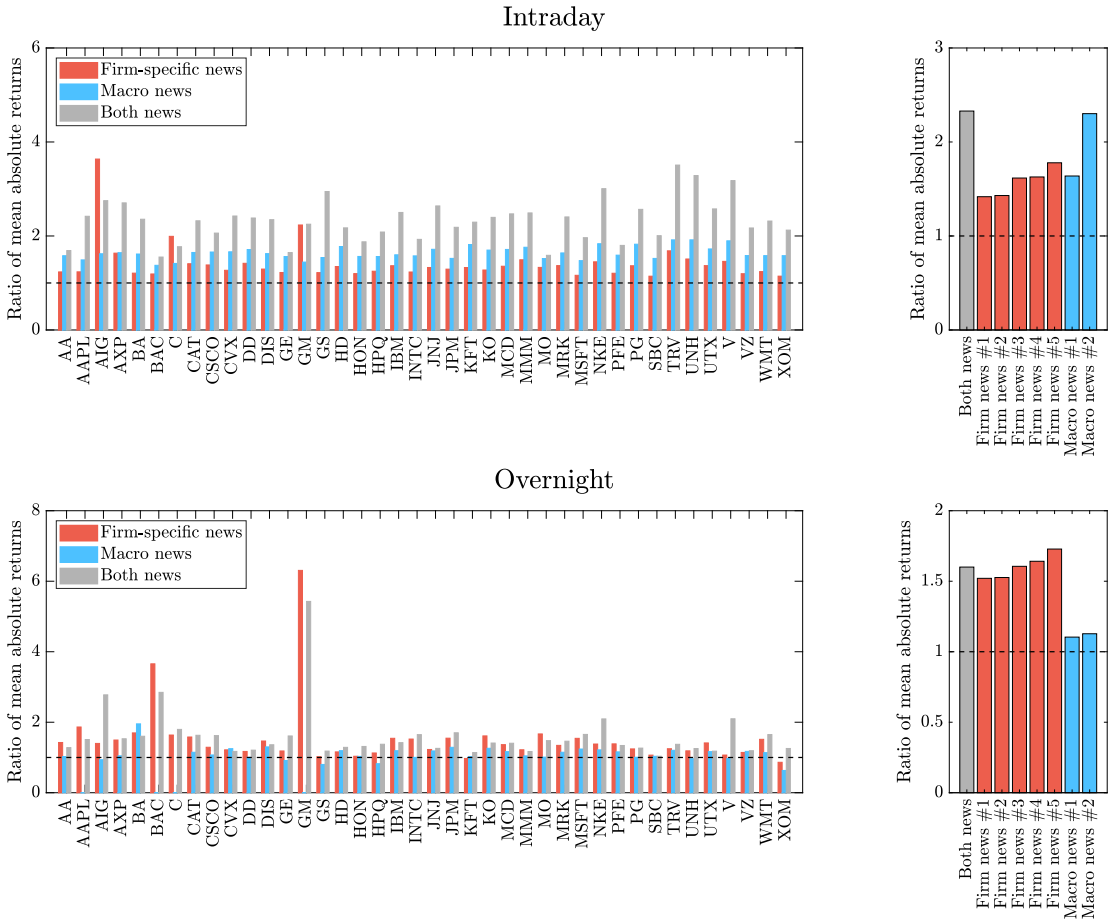

Figure 2: Ratio of mean absolute returns

Note: For the three groups of firm-specific, macro, and both news, the ratio is defined as the mean absolute return for time windows with news divided by the its counterpart without news. The term "return" refers to the abnormal return $\epsilon_{[t_{i-1},t_i]}$, the fitted return $R^{\text{fitted}}_{[t_{i-},t_i]}$, and the total return $R_{[t_{i-},t_i]}$, in the three cases respectively. In the upper left panel, we compare for each company the intraday ratios that are associated with firm-specific news (red), macro news (blue), and both types of news (grey), respectively. The upper right panel aggregates the mean across companies. The bars labelled with "Both", "Firm news #1", and "Macro news #1" represent the means of the grey, red, and blue bars in the upper left panel, respectively. The other red bars (resp. blue bar) are (resp. is) produced in the same way as the one labelled with "Firm news #1" (resp. Macro news #1), except for corresponding to more stringent news filter(s). In the bottom two panels, the computation of overnight fitted and abnormal returns follows (2)–(3), except for changing $R_{[t_{i-1},t_i]}$ (resp. $X^{(k)}_{t_i} - X^{(k)}_{t_{i-1}}$) to the corresponding overnight stock (resp. factor) return, and choosing $\hat{\beta}^C_{t,k}$ and $\hat{\beta}^J_{s,k}$ the same as those from the previous trading day. In the lower left panel, the blue bars for AAPL, BAC, C, and GM are not included, since for each of these four stocks, the number of overnight periods that are associated with macro news exclusively is less than 20.

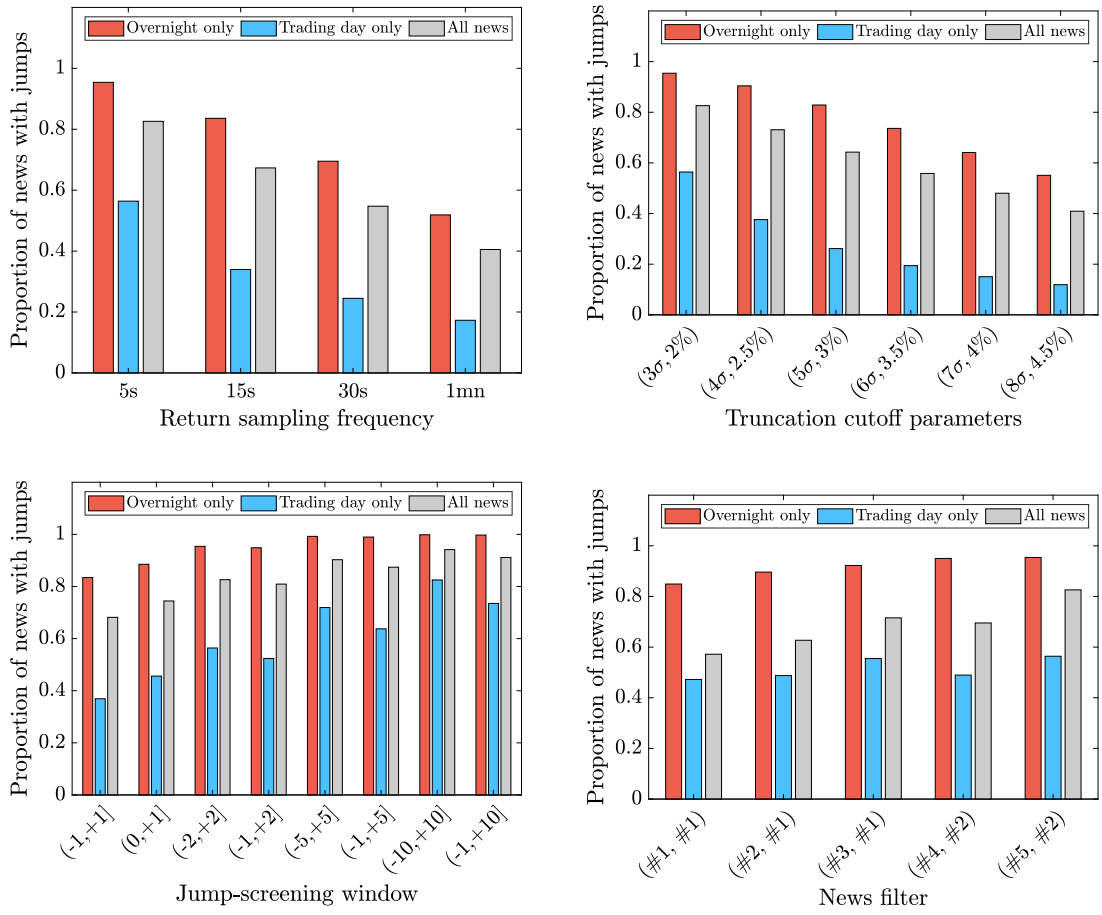

Figure 3: Proportion of news with jumps under various settings

Note: The upper left, upper right, lower left, and lower right panels report the sensitivity of the proportion of jumps led by news to the return sampling frequency, truncation cutoff parameters, news-screening window, and news filter, respectively, while fixing the other parameters at their baseline setting, where we set sampling frequency at 5s, truncation cutoff parameter as 3σ (resp. 2%) for intraday (resp. overnight) returns, jump-screening window as $(t_{i-1} - 2\text{mn}, t_i + 2\text{mn}]$, and firm-specific (resp. macro) news filter as filter #5 (resp. #2). We exclude superseded news items and say a news item is followed by jumps, if the number of jumps led by the news (computed according to the note of Table 4) is greater than zero. The red (resp. blue) bars include only overnight (resp. intraday) news. The grey bars include both. In the upper right panel, the first (resp. second) entry in the pair of truncation parameters from the group label represents the multiple of standard deviations c (resp. absolute threshold c') for detecting intraday (resp. overnight) jumps. The jump-screening windows in the lower left panel are expressed in mn: $(-a, +b]$ indicates that the jump-screening window consists of up to $a$mn before and $b$mn after the 5s interval containing the news. The first (resp. second) entry in each pair of news filters in the lower right panel represents the choice of firm-specific (resp. macro) news filter defined in Table 2.

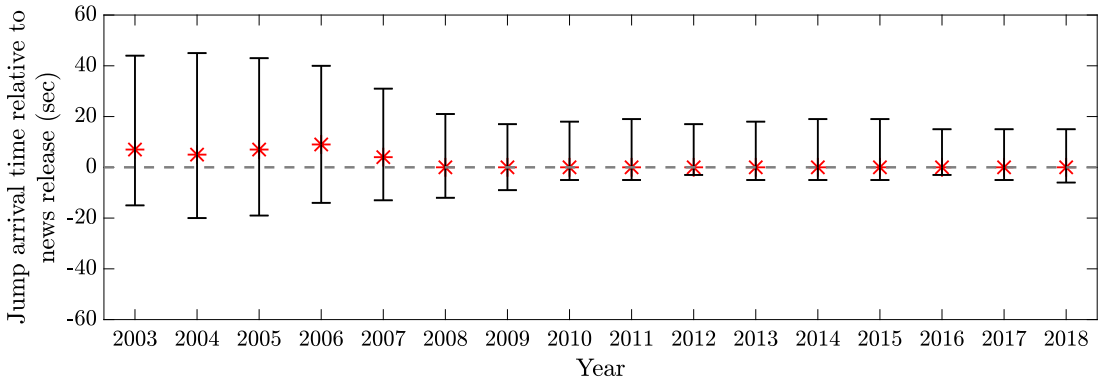

Figure 4: Distribution of jump arrival time relative to news release

Note: This figure shows the evolution over time of the distribution of response time of jumps to news, defined in (5). For each year, the red star, upper and lower edges of the whiskers represent the median, third and first quartiles, respectively. The grey dashed line highlights the level of zero.

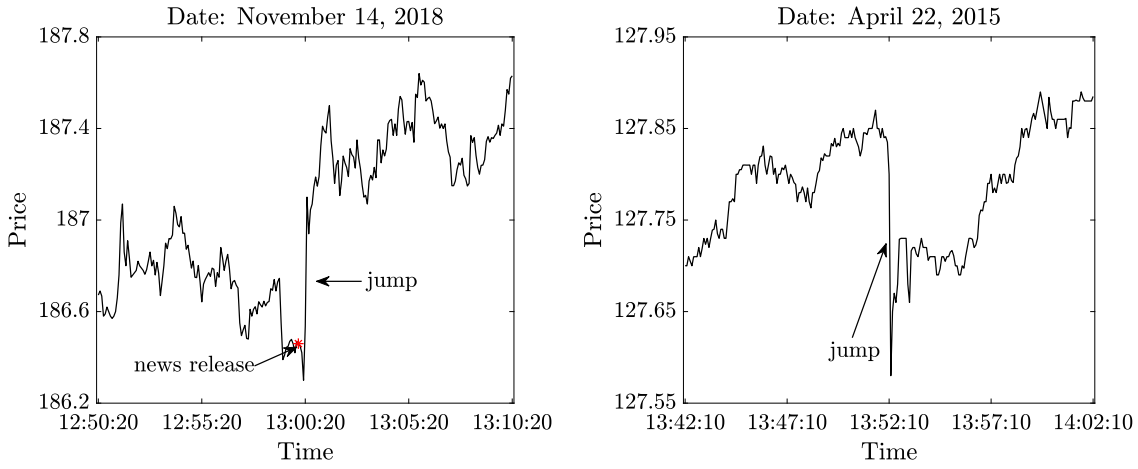

Figure 5: Examples: Jumps with and without news

Note: The left (resp. right) panel shows a representative example of a price jump with (resp. without) news for Apple Inc. on November 14, 2018 (resp. April 22, 2015) at the 5 second frequency. The jump size is 11.1 (resp. 16.6) standard deviations in the left (resp. right) panel. The jump in the left panel follows a news item released at 13:00:00 with the headline "Dialog Semiconductor <DLGS.DE> CEO says not seeing hit to demand from Apple <AAPL.O>". The jump happens 20 seconds later at 13:00:20. The jump in the right panel is not associated with any news in the database, even using the widest news filters #1 for both firm-specific and macro news, and searching for news in the long window $[t_{i-1} - 10\text{mn}, t_i + 10\text{mn})$ surrounding the jump time.

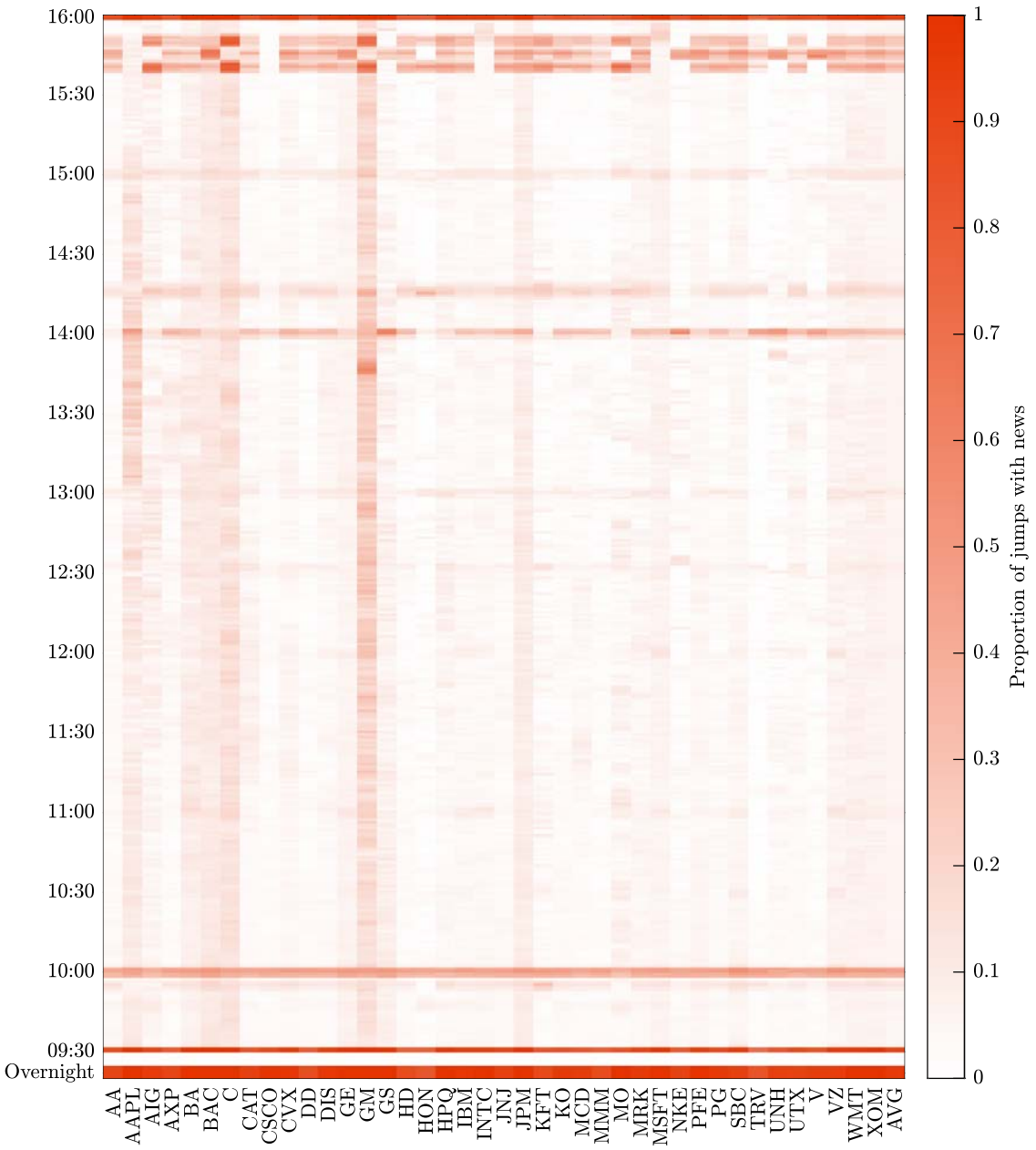

Figure 6: Proportion of jumps with news by stock and hour

Note: For each company and for each overnight interval (lowest row) or 1mn interval between 09:30 and 16:00, we count the numbers of jumps with and without news in each interval across the whole sample period, and then compute the proportion of jumps with news in the news-screening window. The column "AVG" at the right end reports the average across all stocks.

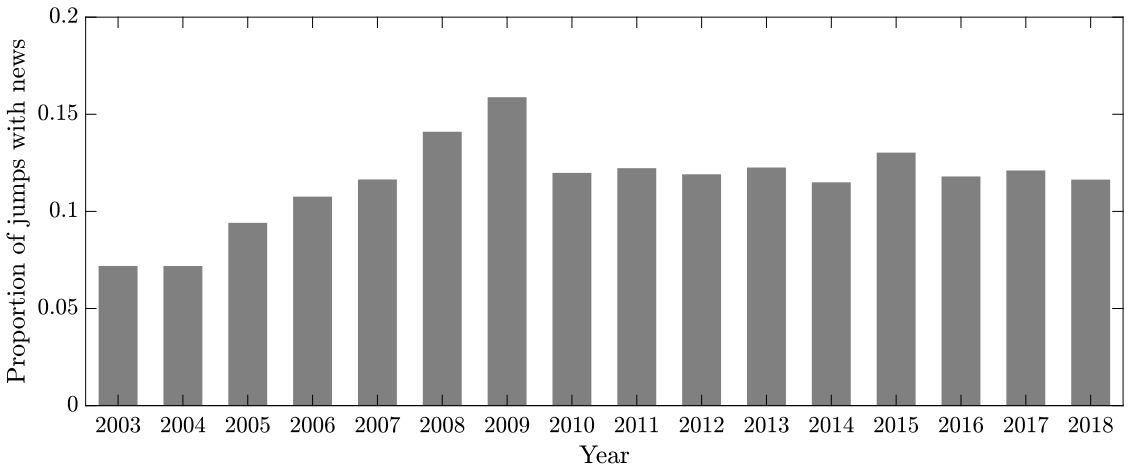

Figure 7: Proportion of jumps with news over time

Note: This figure shows the proportion of jumps with news over time, computed year by year across stocks, including both intraday and overnight jumps. Jumps are considered associated with news if at least one idiosyncratic news item under filter #1 or one macro news item, also under filter #1, is present in the jump's news-screening window.

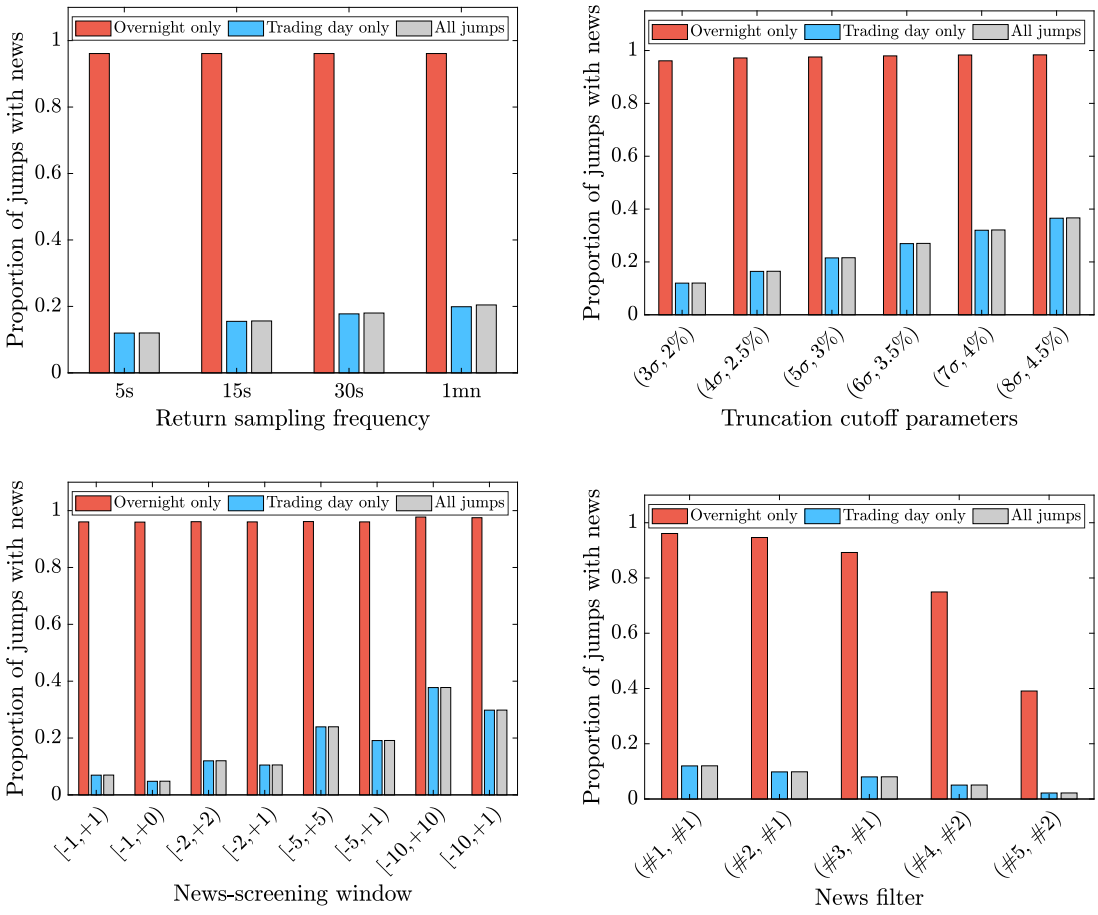

Figure 8: Robustness on the proportion of jumps with news

Note: Symmetric to the four panels in Figure 3, this figure shows the sensitivity of the proportion of jumps with news under various settings. The settings for producing these four panels are the same as those for producing Figure 3, except for the following four changes. First, we change $\mathbb{P}(\text{jump}|\text{news})$ to $\mathbb{P}(\text{news}|\text{jump})$. Second, we replace the jump-screening window to news-screening window in the lower left panel. Third, the bars labelled with "Overnight only" (resp. "Trading day only") represent the jumps whose news-screening windows cover (resp. do not cover) overnight periods, including jumps before 09:32, after 15:58, and overnight jumps (resp. jumps between 09:32 and 15:58 during regular trading sessions). Fourth, we set the baseline firm-specific (resp. macro) news filter as filter #1 (resp. #1), so as to show the intraday and overall proportions of jumps with news are low, even if one takes the full universe of news into consideration.

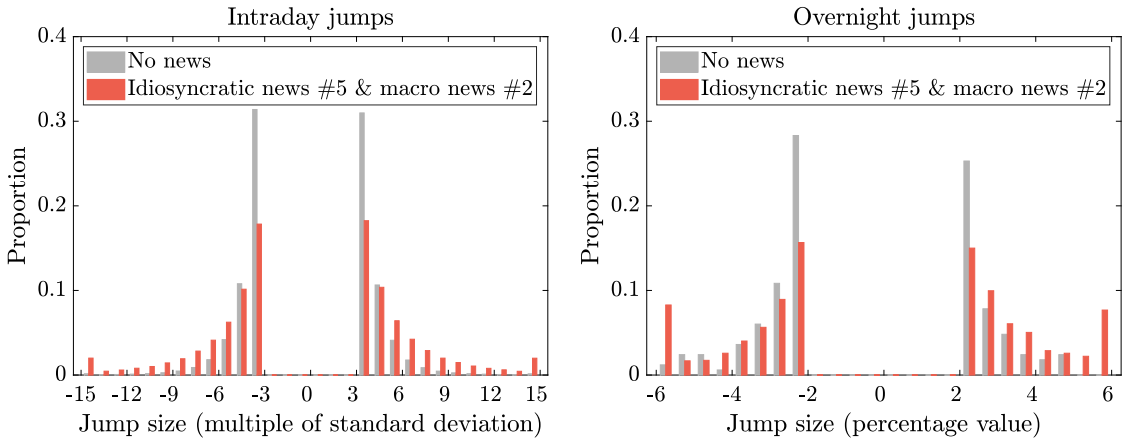

Figure 9: Distribution of jump sizes with and without news

Note: The left (resp. right) panel shows the distribution of intraday (resp. overnight) jump sizes across all stocks and dates. In each panel, we compare the distribution of jump sizes for jumps without news (in grey) and for jumps driven by idiosyncratic news under filter #5 or macro news under filter #2 (in red). The histograms are truncated at intraday jump sizes greater than 15 standard deviations and overnight jump sizes greater than 6%.

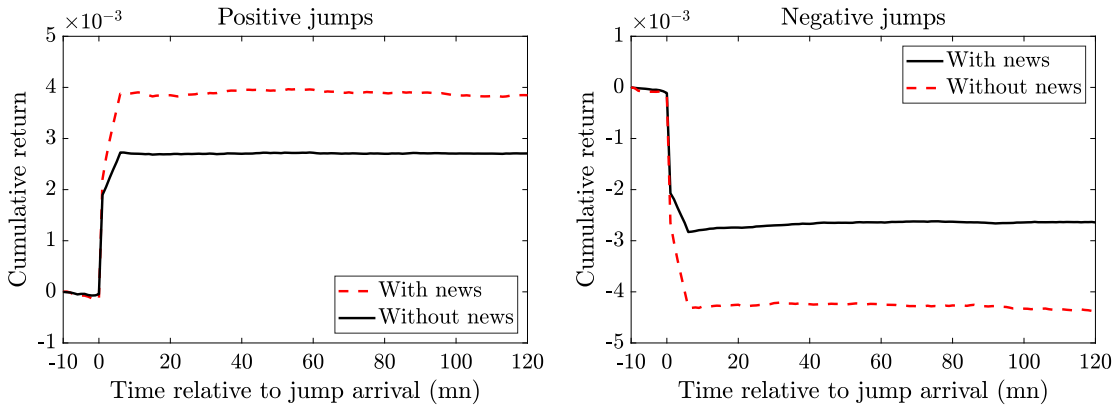

Figure 10: Persistence of the price impact of intraday jumps with and without news

Note: The left (resp. right) panel shows the changes of cumulative return (CR) CR$[-10, b]$ around positive (resp. negative) intraday jumps at the 1mn frequency. To compute the CR, we begin with aggregating adjacent jumps into a cluster, if the time gap between any two successive jumps is less than 5mn. For each cluster of jumps, we treat the arrival time of the first jump t_{i-1} as the origin of time in both panels. The direction of each cluster of jumps is said to be positive (resp. negative) if its aggregate impact within the first 5mn CR$[0, 5]$ is greater (resp. less) than zero, and the cluster of jump is said to be associated with news if at least one jump in the cluster is associated with news. We then compute the log-returns $R_{[t_{i-1}+k, t_i+k]}$ for all integers $k \in [-10, 120]$. Returns are regarded as missing if either $t_{i-1}+k$ or t_i+k is outside the regular trading hours. Finally, for each 1mn interval, we compute the mean of all non-missing returns and CR is the partial sum of these averaged returns.

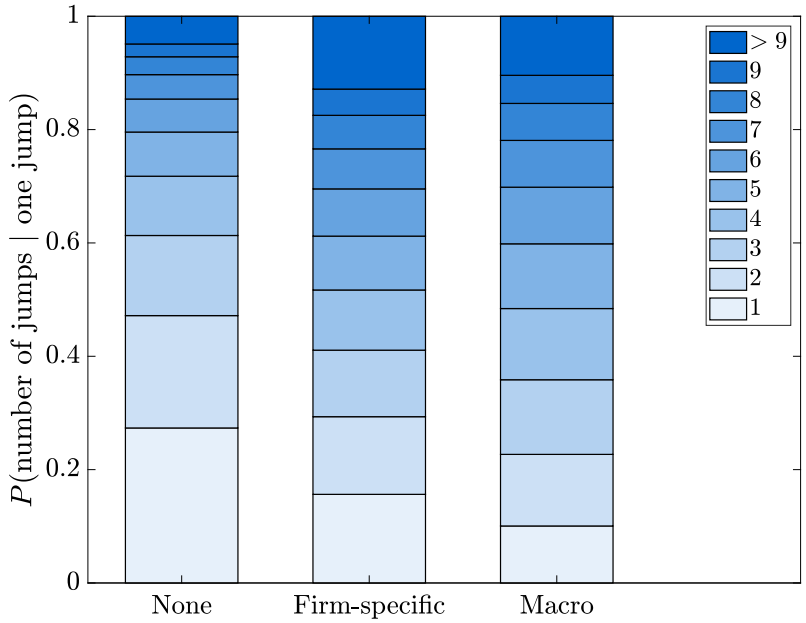

Figure 11: Clustering of jumps

Note: The three bars labelled with "None", "Firm-specific", and "Macro" show the cumulative probability of different numbers of clustered jumps within a time window, given one observes in $[t_{i-1}, t_i)$ a jump without news, as well as a jump driven by firm- or industry-specific news and macro news, respectively. If there exist both firm-specific and macro news items within the news-screening window $[t_{i-1} - 2\text{mn}, t_i + 2\text{mn})$ of a jump, we determine the type of news driving this jump case-by-case: (1) Within the news-screening window, suppose there exists at least one macro news item from macro news filter #2. Then, the jump is said to be driven by macro news (resp. both macro and firm-specific news), if no (resp. at least one) firm-specific news item survives from firm-specific news filter #5; (2) suppose there is no macro news item from macro news filter #2, but at least one news item from macro news filter #1. The jump is said to be driven by macro (resp. firm-specific) news if there is no (resp. at least one) firm-specific news item survives from firm-specific news filter #3 (resp. #4). The jump is said to be driven by both macro and firm-specific news, if at least one firm-specific news item survives from firm-specific news filter #3, but no news survives from firm-specific news filter #4. A jump is repeatedly included in both two bars labelled with "Firm-specific" and "Macro", if it is driven by both firm-specific and macro news. In addition to the baseline setting for jump detection, we set the time window for counting clustered jumps the same as the news-screening window for simplicity. In other words, for each given jump at $[t_{i-1}, t_i)$, we count the number of consecutive jumps that center around the given jump within $[t_{i-1} - 2\text{mn}, t_i + 2\text{mn})$ and are driven by the same type of news as the given jump.

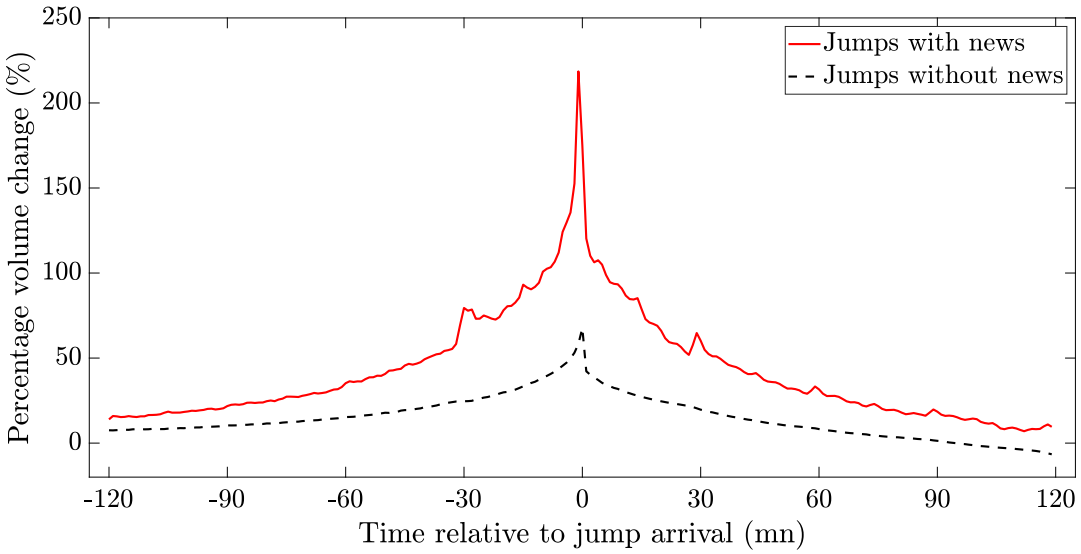

Figure 12: Percentage change of volume around jumps

Note: For each stock and each intraday jump arriving at $[t, t+5\text{s})$, we compute the percentage volume change of the stock in each 5s interval between $t-2$ hours to $t+2$ hours. Here, the percentage volume change is defined as the difference between the volume in the 5s interval and the referenced volume divided by the referenced volume, where the referenced volume is the average volumes of the stock across all 5s intervals on the last trading day. To smooth the curves, for each 1mn interval between $t-2$ hours to $t+2$ hours, we compute the moving average of twelve 5s percentage volume changes as the percentage volume change per 5s for the corresponding 1mn interval. Finally, each point on the red solid (resp. black dashed) curve averages across all jumps with (resp. without) news for all stocks.

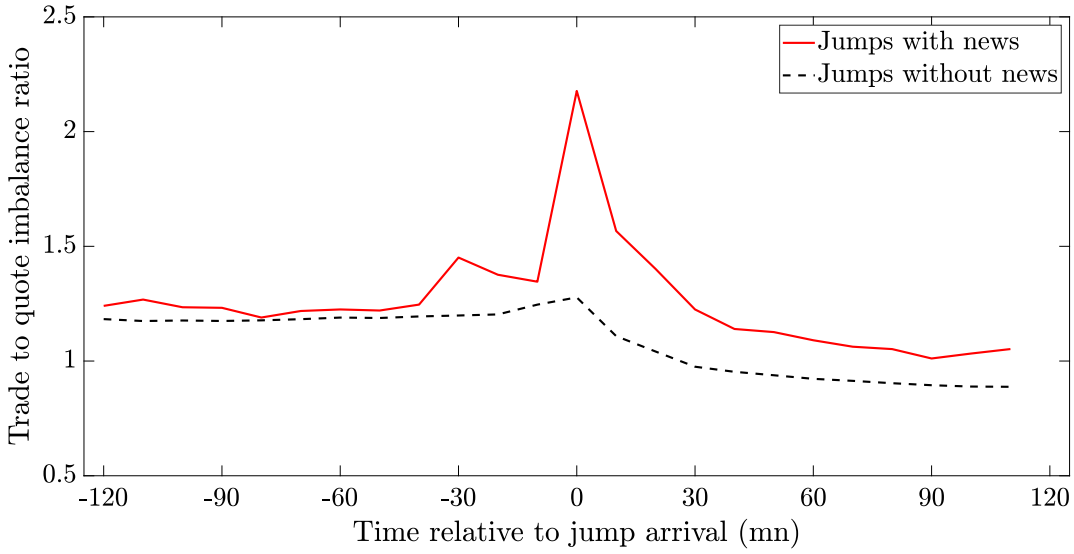

Figure 13: Trade-to-quote imbalance ratio around jumps

Note: For each stock and each intraday jump arriving at $[t, t+5\text{s})$, we compute the trade-to-quote imbalance around jumps in each 5s interval between $t-2$ hours to $t+2$ hours. Here, for each 5s window, the trade-to-quote imbalance is computed case-by-case. If the 5s log-return is a positive (resp. negative) jump, the trade-to-quote imbalance is defined by the buy-side (resp. sell) imbalance, which is total buy (resp. sell) size divided by the total best ask (resp. bid) size. If the 5s log-return is not detected as a jump, the trade-to-quote imbalance is defined by the average of the buy-side and sell-side imbalance. For each jump, we then compute the trade-to-quote imbalance ratio as the ratio between the trade-to-quote imbalance in the 5s interval and its referenced counterpart, where the referenced counterpart is the average trade-to-quote imbalance of the stock across all 5s intervals without jumps on the last trading day. To exclude extreme ratios, we smooth the curve within 10mn rolling windows, similar to Figure 12. Finally, each point on the red solid (resp. black dashed) curve represents the average across all jumps with (resp. without) news for all stocks.

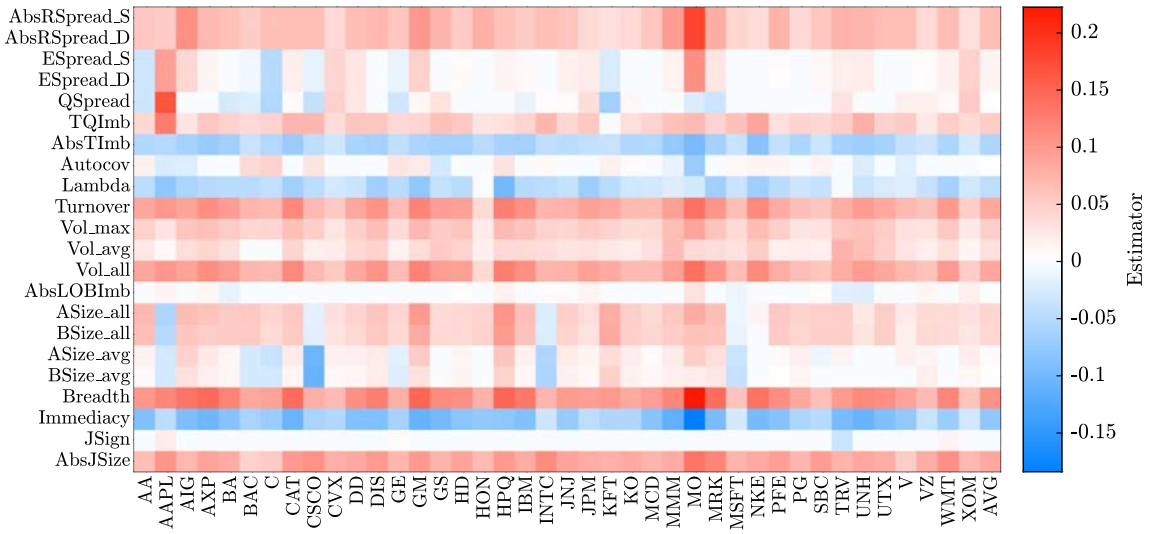

Figure 14: Estimation results of regressions (8) under the baseline setting

Note: Columns "AA"–"XOM" represent the estimation results of regressions (8) for individual stocks. We report in each cell the corresponding estimator of β. We set an estimator as zero if it is insignificant at the level of 5%. Each cell in Column "AVG" averages all preceding estimators in the same row.