Exogenous and Endogenous Price Jumps Belong to Different Dynamical Classes

Riccardo Marcaccioli,[1,2] Jean-Philippe Bouchaud,[1,3,4] and Michael Benzaquen[1,2,3,*]

[1] *Chair of Econophysics and Complex Systems, Ecole polytechnique, 91128 Palaiseau Cedex, France*
[2] *LadHyX UMR CNRS 7646, Ecole polytechnique, 91128 Palaiseau Cedex, France*
[3] *Capital Fund Management, 23-25, Rue de l'Université 75007 Paris, France*
[4] *Académie des Sciences, Quai de Conti, 75006 Paris, France*

Synchronising a database of stock specific news with 5 years worth of order book data on 300 stocks, we show that abnormal price movements following news releases (exogenous) exhibit markedly different dynamical features from those arising spontaneously (endogenous). On average, large volatility fluctuations induced by exogenous events occur abruptly and are followed by a decaying power-law relaxation, while endogenous price jumps are characterized by progressively accelerating growth of volatility, also followed by a power-law relaxation, but slower than for exogenous jumps. Remarkably, our results are reminiscent of what is observed in different contexts, namely Amazon book sales and YouTube views. Finally, we show that fitting power-laws to *individual* volatility profiles allows one to classify large events into endogenous and exogenous dynamical classes, without relying on the news feed.

I. INTRODUCTION

Earthquakes, disease outbreaks, volcanic eruptions, avalanches, species extinctions, traffic jams, economic crises and financial crashes are but a few examples of a long list of extreme events that upend natural and social systems. Given their ubiquitous presence (and their relevance in our everyday life), they have received a great amount of attention from different scientific communities [1–4]. A central question that researchers have tried to answer is whether these events are caused by *exogenous* events (like the meteorite which probably triggered the Cretaceous–Paleogene extinction event) or result from some amplifying feedback mechanism internal to the system, in which case the shock is *endogenous* [5].

This topic is particularly important in the context of financial markets, and is related to the long-standing Efficient Market controversy. If markets are efficient, significant price movements can only be due to unpredictable exogenous shocks. On the other hand, if self-reflexive feedback loops are present, extreme price displacements can be triggered by small (and seemingly irrelevant) fluctuations, which can ultimately generate substantial excess volatility.

Is it really possible to categorize extreme events into exogenous and endogenous? Answering this question in a general context is highly non-trivial [5]. Nevertheless, building on the idea that endogenous shocks can only appear in systems that are somehow "fragile", i.e. close to an instability, a methodology to differentiate exogenous from endogenous events in empirical data has been proposed in a very interesting series of papers [6]. Hawkes processes, in particular, provide a convenient and versatile modelling framework consistent with the assumption of a near-critical system. In fact, Hawkes processes were introduced to model self-exciting earthquakes [7, 8], but have been shown to be relevant in many other contexts, such as financial market activity [9], crime outbursts [10], or banking and corporate defaults [11], to name a few.

The proximity of an instability suggests the use of *critical* Hawkes processes, characterised by a power-law memory kernel. Such a specification has allowed the authors of [12] to efficiently discriminate endogenous from exogenous bursts of views among 5 millions YouTube videos. Views spikes which result from a contagion process on the underlying social network of influence are characterised by a slow (power-law) post-shock relaxation to the baseline views' number distribution and by an *almost* mirror-image pre-shock growth. On the other hand, spikes that are chiefly induced by exogenous shocks are characterised by a faster post-shock decay, almost without any pre-shock growth (see Fig. 1 for an illustration in the case of market price jumps). Quite remarkably, similar results also hold for a dataset of Amazon books sales [13, 14].

In the context of financial price time series, such alluring findings are still lacking. Nevertheless, several past studies hint at the possibility of dividing exogenous and endogenous extreme events in a similar manner. The observable of interest of most of these studies is the instantaneous volatility profile of stocks (log-)returns. First of all, it has been shown in [15–20] that the rate of abnormally large absolute returns after a large exogenous shock decays as a power-law (which corresponds to the so-called Omori law in the context of earthquake aftershocks [21]). A similar behaviour has also been observed in the absolute value of the returns [15, 22] (sometimes restricting to abnormally high returns [19, 23, 24]) and in the dynamics of the bid-ask spread [20, 25]. Quite strikingly, these findings seem to hold independently of the considered type of exogenous event, asset type or timescale over which returns are computed and therefore hint to the existence of a possible universality class. This is the "Efficient Market Class" (EMC), in the sense that the market price strongly reacts to unexpected large external shocks. Such strong reactions to exogenous events are not only limited to prices and are indeed known to

*Electronic address: michael.benzaquen@polytechnique.edu

be present in market activity as well [26, 27].

Within the Efficient Market picture, all extreme events are exogenous, and no endogenous "Self-Exciting Class" (SEC) should be detectable. However, this view has long been challenged by a series of investigations, starting with the seminal work of Cutler, Poterba and Summers in 1989 [28] where they conclude that the *evidence that large market moves often occur on days without identifiable major news releases casts doubt on the view that stock price movements are fully explicable by news.* This conclusion, drawn using daily returns, was confirmed in later studies [29, 30] looking at different time granularity. In particular, the authors of Ref. [31] found that most large intra-day price fluctuations for single stocks (called "jumps" henceforth) happen independently from news releases.

These latter findings point to the possibility that a non-efficient, SEC exists and that it can be successfully identified in empirical data, with jumps induced by a direct feedback loop between trades and volatility. To expose the existence of such endogenous jumps, some studies take a granular approach and calibrate a Hawkes process on trade-by-trade data [32–35] while others focus on macroscopic quantities like volatility, volumes or trends [36–38].

In this study we follow this latter stream of literature. In particular the present work builds upon the observations presented in Ref. [31]. By using a database of news concerning single or multiple stocks, it has been empirically established that the average volatility profile following a news-induced, EMC jump is markedly different from the one following an SEC jump.

We broadly confirm and significantly extend the results of Ref. [31], using a more recent database (300 different stocks traded at the NYSE from 01/01/2015 to 01/01/2020). We show that it is indeed possible to classify extreme price moves into two distinct dynamical classes, EMC and SEC. Inspired by seismology studies, we argue that instead of individual price jumps and individual news one needs to focus on clusters of jumps and clusters of news. We show how SEC clusters of jumps (not triggered by news) display very different properties from EMC clusters (closely following a cluster of news). Consistent with previous studies on YouTube views and Amazon book sales [12, 13], we find that the average volatility profile of SEC events is much more symmetric than the average profile of EMC events and that, while they both decay as power-laws, they are characterised by different relaxation exponents. In fact, the observed values of these exponents are remarkably close to those reported for YouTube views and Amazon book sales [12, 13]. In agreement with the recently introduced endogenous liquidity crises model [38], SEC cluster of jumps appear to be preceded by a slow increase in volatility and price trends.

Having established the *average* profile of EMC and SEC jumps, we then turn to analyzing *individual* volatility profiles around large clusters of jumps. Determining the shape (parameterized by a power-law before and after the first jump of a cluster) and asymmetry of these profiles allows us to classify jumps into EMC and SEC types with remarkably high degree of success, as measured by the Area Under Curve of the corresponding classification tasks. Finally, we discuss several wider implications of our findings.

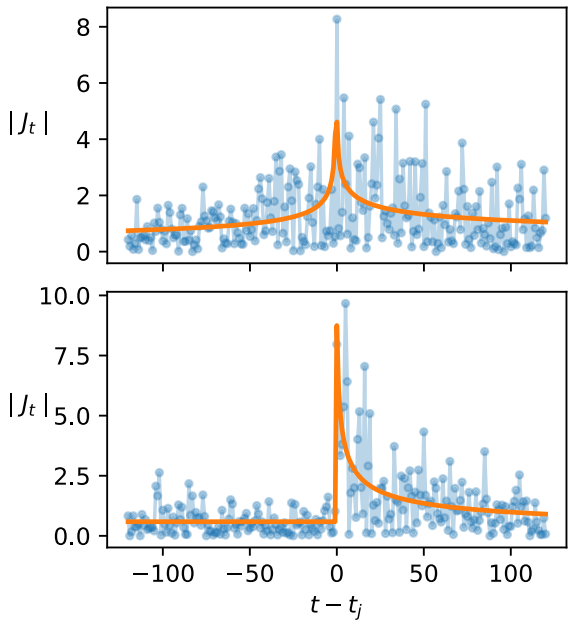

FIG. 1: Examples of endogenous (top) and exogenous (down) bursts of volatility. We draw in orange the best fits using the functional form given by Eq. 3 below. Exogenous shocks are characterised by a slow power-law precursory growth and an almost symmetric relaxation. Endogenous shocks are asymmetric around the instant of the shock and display a faster relaxation toward the pre-shock activity levels.

II. EXTREME EVENTS IN ELECTRONIC MARKETS

A. The Limit Order Book

In the present days, most of the world's financial markets use an electronic trading mechanism called a Limit Order Book (LOB) to facilitate trade of a given asset. The LOB $\mathcal{L}(t)$ is the collection of all active limit orders at any given time t and, as such, it can be thought as a concise representation of the supply and demand of any electronically tradable asset. The LOB is usually divided into the ask side (the set of active sell limit orders) and the bid side (the set of active buy limit orders). The highest (lowest) occupied buy (sell) price level is called the ask a_t (bid b_t). The difference between the two is called the spread $s_t = a_t - b_t$ while their average $m_t = (a_t + b_t)/2$ is called the mid-price and it is usually

used as a proxy for the price of an asset (see [39] for more on this topic). Each price level at or below the bid (above the ask) is populated by volumes $V_b, V_{b-1}, \dots V_{b-n}, \dots$ $(V_a, V_{a+1}, \dots V_{a+n}, \dots)$. This means in particular that a buy market order of size $Q_n = V_a + V_{a+1} + \dots + V_{a+n}$ leads to an immediate ask price move up by n ticks. (Note that some price levels maybe empty). The quantity n/Q_n can be thought of as a measure of sparsity of the LOB on the ask side, with a similar definition for the bid side. We refer the interested reader to Ref. [40–42] for extensive reviews on the empirical properties of LOBs.

B. Dataset description

1. *Order book data*

We conduct the analysis by using the four best price levels (2 for the bid and 2 for the ask) from the LOB of a selection of 300 stocks continuously traded on the NYSE from 01/01/2015 to the 01/01/2020. Each LOB is sampled on a minute timescale. These snapshots portray the time evolution of the price and the supply and demand of a given stock. We only consider data collected during the regular US trading session (which start at the 9:30 a.m. and ends at the 4:00 p.m.) and only those sessions with a moderate or high trading activity (we only keep trading days with at least 300 recorded price changes). The reasons for this latter filtering step are threefold. First of all, to uniformise our sample: some stocks are always characterised by a continuous moderate or high trading activity while others are not. Secondly, to avoid spurious effects: our jump detection methodology assumes that the returns (after standardization) are approximately distributed as a standard normal; this assumption crumbles when the market activity is low, leading to the detection of numerous spurious jumps [43]. Lastly, estimation accuracy: estimating scaling laws is a notoriously hard task [12], especially in highly noisy environments [44]. We noted that removing from our samples those days with an exceedingly high amount of zeros or missing values in the volatility series led to a more accurate estimation process. Coherently with the moderate activity requirement, we selected the stocks based on their 2019 turnover. For a complete list of all the stocks included in our analysis, as well as a detailed description of the data pre-processing, see the Appendix.

2. *News data*

We use a generic (i.e. not only finance related) news database which contains articles published on Bloomberg during the same period (01/01/2015 to 01/01/2020). Each news item is characterised by its title, the time at which it was been posted online and a list of tickers (i.e. unique stocks identifiers) which the news may concern. For this study we will only consider those news which are marked as relevant for at least one of the 300 stocks we consider and which explicitly display in the title the ticker of

How to get the dataset with the related ticker list?

1. at least one of the stocks it may concern, or
2. at least one of their companies' names, or
3. at least one of their companies' abbreviated names (i.e. IBM instead of International Business Machines, or Abbott instead of Abbott Laboratories).

See the Appendix for summary statistics of the news, as well as their distribution across times and stocks.

C. Price Jumps Detection

The observable we shall focus on in our analysis is the mid-price m_t. Before exposing possible differences between exogenous and endogenous extreme mid-price movements, we need a way to assess which variations $m_t - m_{t-1}$ can be considered extreme or abnormal. In order to do so, we follow the non-parametric price jumps detection methodology proposed in Ref. [45] and further refined in Ref. [46]. The intuition behind such procedure is very straightforward: fluctuations of the mid-price m_t are first normalized so that, in the absence of jumps, their distribution is as close as possible to a standard normal distribution. Once this normalization is properly defined, Extreme Value Theory can be used to derive a threshold above which a fluctuation can be classified as a jump within a given probability level.

We consider the 1-minute return time series $r_t = \log \frac{m_t}{m_{t-1}}$. Mid-price returns are known to have approximately zero mean but a strongly fluctuating variance, with both intra-day seasonalities and long-memory, intermittent dynamics (see e.g. [33, 37, 47–49]). As such, any standardization procedure must take into consideration both the instantaneous evolution of the variance as well as any possible seasonality. We therefore define the "jump-scores" J as:

$$J_t = \frac{r_t}{\sigma_t f_t}, \tag{1}$$

Is the rolling window across multiple days?

where $\sigma_t^2 = \frac{\pi}{2K} \sum_{i=1}^{K} |r_{t-i}||r_{t-i+1}|$ is an estimator of the local volatility over a rolling time window of length $K = 390$ (i.e. one day worth of data, but dropping any overnight contribution) and f_t is an estimator of the intraday periodicity component (see the Appendix for its detailed definition).

Under the null hypothesis of no jumps and a vanishing sampling frequency, the statistics of the maximum of $|J_t|$ converges to a Gumbel distribution. One can therefore reject, with a statistical significance $\alpha = 0.01$, the null hypothesis of absence of jumps whenever we observe:

The max of every jump? or the max jump score every day?

$$|J_t| > C_K - S_K \log\left(\log \frac{1}{1-\alpha}\right) \approx 4.36, \quad (K = 390).$$

The constants $S_K = (2\log K)^{-0.5}$ and $C_K = (2\log K)^{0.5} - (\log\pi + \log(\log K))/(2(2\log K)^{0.5})$ are dependent on the window size and are meant to correct for the fact that, within each window, we are performing multiple hypothesis tests. As such, by using this threshold, one expects to find only α spurious jumps in a given sample of K observations.

In a nutshell, we mark as "jumps" those price movements with associated z-scores that are approximately 4-sigma away from zero. In order to avoid effects due to market opening or closing, we discard jumps happening in the first or last 15 minutes of the trading day.

D. Clusters of jumps

We run the price jumps detection methodology outlined above on all the 300 mid-prices time series. We record a total of 258,671 jumps. The daily average number of jumps of a given stock ranges from 0.25 to 3.26, with an average value of 0.70 and a standard deviation of 0.42.

As such, on average, we would expect a stock to jump about once per day (which is, in passing, much more frequent that the expected number of news that can shake the value of a given stock). However, if we look at the inter-time distribution between two consecutive jumps within the same day (Fig. 2a)) we observe clear deviations from a Poisson law. Rather, the distribution is well fitted by a power-law behaviour. Power law distribution of waiting times is a typical fingerprint of many social activities [50], including trading in financial markets, but also seismic activity or epileptic activity, see e.g. [51, 52]. Such "bursty" time series are often modelled in terms of self-exciting Hawkes-like point processes [9]. Indeed, our empirical findings are consistent with previous observations, see e.g. [43].

Such long-memory effects in the dynamics of jumps can potentially induce spurious effects when an aggregate analysis is performed. Consequently, and following common practice in seismology [8], we move away from earlier studies on price jumps [31, 53] and instead of analysing single jumps, we focus on *clusters* of jumps.

We adopt a simple and intuitive clustering technique to group jumps together: we compare the observed inter-times between any two consecutive jumps against the one prescribed by a Bernouilli null-hypothesis, corresponding to independent jumps occurring with probability p. If, under the null hypothesis, the probability of observing the given inter-time is smaller than a significance level ϵ, we cluster the two jumps together.

It is straightforward to show that, under this simple null model, given the presence of a jump at time t_0, a second jump occurring at time t_1 is assigned to the same cluster when:

$$t_1 - t_0 < \frac{\log(1-\epsilon)}{\log(1-p)} - 1\,. \tag{2}$$

We set $\epsilon = 0.05$ and we determine p in order to have our null model preserving, on average, the number of jumps of each stock within any given month.

After running our clustering methodology, we find a total of 197,197 clusters of jumps (most made out of one single jump). The daily average number of clusters of jumps for a given stock ranges from 0.17 to 2.77 with an average value of 0.53 and a standard deviation of 0.37. The normalized inter-time distribution of those clusters happening in the same day is now well described by an exponential distribution (Fig. 2a). This feature validates that such clusters can be reasonably considered to be independent, and therefore that spurious effects induced by any aggregation procedure are reasonably reduced.

E. News Related Jumps

Similarly to jumps, we also observe that news releases tend to cluster in time. We therefore perform on news the same clustering procedure applied to jumps. We then mark as news related (or exogenous) those clusters of jumps which started up to one minute before and up to four minutes after the beginning of a cluster of news. This is done in order to account for the fact that a particular news may have become available to some market participants before our recorded news release timestamp and to account for possible misalignments between the news feed and the LOB data. We mark as not news-related, or endogenous, the remaining clusters of jumps. For simplicity, we exclude from our analysis those endogenous clusters which start within a cluster of news.

Another effect that one should need to consider is the role of macroeconomic news (not specific to a given stock), which might trigger clusters of stock price jumps. A systematic identification of possibly relevant macroeconomic news would entail contextual word recognition and goes beyond the scope of the current work. To circumvent such a limitation, we remove from our list of clusters those which participate in a market-wide or sector-wide event, as any relevant macroeconomic news would trigger. Hence we compute, for each cluster of jumps, the number of stocks which display an overlapping cluster of jumps during the same time interval. Whenever we observe a number of overlapping clusters higher than 30 (10% of the stocks in our pool), we mark them as belonging to a market-wide or sector-wide event. Changing this threshold to 15, 60 or to one prescribed by a null hypothesis of clusters independence (detailed in the Appendix), does not significantly affect our findings.

As a final safe-guard, we also exclude from our samples clusters of jumps, of a given stock, happening within 100 minutes from each other. This is done following Ref. [53] in order to completely avoid any contamination effect that may happen in our analysis.

Finally, we are left with a total of 106,680 clusters of jumps, out of which only 1073 are news related (note that most major, company related news happen outside

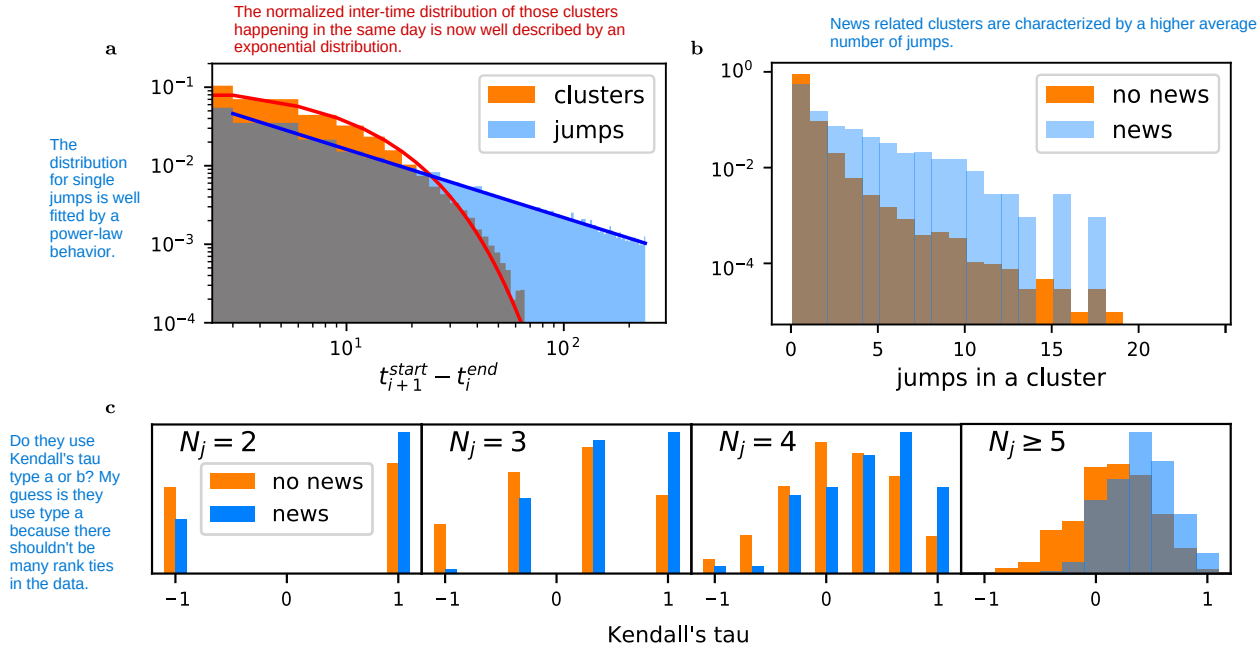

FIG. 2: Characterization of identified clusters of jumps. **(a)** Inter-times distribution of jumps and clusters of jumps. We normalize with the minimum time granularity to detect the given event, i.e one minute for jumps and L minutes (Eq. 2) for clusters of jumps. The blue line is the best fitting power law (exponent 0.88) and the red line is the best fitting exponential (rate 0.11). **(b)** Distribution of the number of jumps within exogenous ("news") and endogenous ("no news") clusters. **(c)** Distributions of Kendall's tau rank correlations between the amplitude ranking of the jumps in a given cluster with N jumps and their chronological ranking.

Kendall's Tau-a is a non-parametric statistical measure used to assess the strength and direction of the ordinal association between two variables. Unlike Kendall's Tau-b, which accounts for tied ranks, Kendall's Tau-a is used when there are no tied ranks in the data.

Advantages of Using Kendall's Tau-a:
- Non-parametric: Does not assume a specific distribution for the data.
- Ordinal Data: Can be used with ordinal, interval, or ratio data.
- Robust: Less sensitive to outliers compared to Pearson's correlation.

market hours).

III. RESULTS

A. The Internal Structure of Clusters

We now compare the composition of endogenous and exogenous clusters. In Fig. 2b, we observe that news-related clusters are characterised by a higher average number of jumps. In Panel Fig. 2c, we compare the distribution of Kendall's tau correlation [54] between the chronological ranking of jumps and the ranking of jumps based on their amplitude. A value $\tau = 1$ corresponds to the case where the jump happening first is also the largest, the second jump is the second highest and so on. A value $\tau = -1$ corresponds to a sequence of jumps happening in "reverse order", the largest one being the last of the cluster. Beside accommodating more jumps, news-related clusters are more naturally ordered in time than endogenous clusters, consistent with the idea that exogenous events are strong and sudden responses to an external shock while endogenous events are the result of a self-exciting stochastic process, with a progressive build-up.

Motivated by this consideration, and by the well-documented assertion that instantaneous mid-price variations can be described by means of Hawkes processes (see Ref. [9] for a recent review), we now show how exogenous (EMC) and endogenous (SEC) events are characterised by markedly different profiles of instantaneous volatility, price trend and LOB sparsity.

B. Average Profile of EMC and SEC Jumps

In order to show that clusters of jumps which are triggered by a news release display markedly different characteristics from those which are not, we focus on the following five quantities:

- Instantaneous jump-score: $|J_t|$;
- Exponential moving average of past excess volatility, defined as:

$$\Sigma_t = \kappa |J_t| + (1-\kappa)\Sigma_{t-1}\,,$$

where κ defines the averaging timescale, here chosen to be $\kappa = 0.12$ (corresponding to a decay time of 16 minutes, see [37, 38]). Note that we exclude from the exponential average calculations the standardized returns J_t marked as a jumps;

- Normalized past price trend:

$$T_t = \kappa J_t + (1-\kappa)T_{t-1}\,,$$

using the same value of κ as above, and the same exclusion of jumps in the computation;

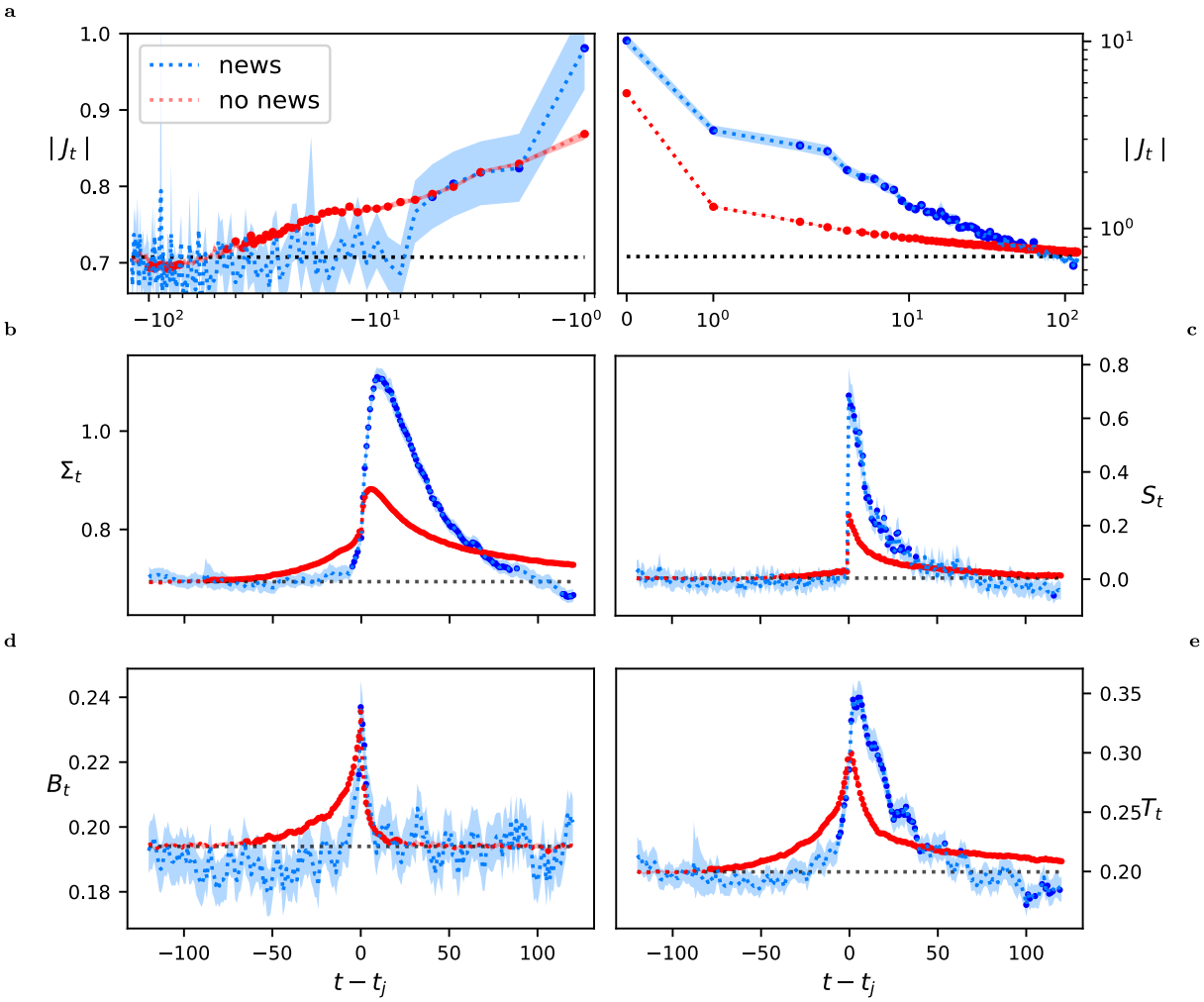

FIG. 3: Differences between clusters of jumps which happened in close proximity to a news release and those which did not. **(a)** In blue we show the average of the absolute jump-score $|J_t|$ only when news related clusters are considered, in red we highlight the average across the remaining clusters. The 0.01 confidence bands (lighter colours) on those averages are obtained using bootstrapped samples. For each point in time and for each of the two sub-samples, we perform a Welch's t-test against the distribution of $|J_t|$ coming from the first 20 minutes of our observation window. We use a marked dot whenever we can reject, at 0.01 significance, the null hypothesis that the two distributions have the same mean. We apply the FDR method to account for the multiple tests performed. Note that the power of the test is different between the news and no news case given their different sample sizes. We use a log scale to highlight the power law behaviour of the instantaneous volatility. **(b)** Same plot for past excess volatility Σ_t, in linear scale. **(c)** Same plot for the instantaneous normalized LOB sparsity S_t, in linear scale. Notice the small decrease of liquidity starting 15 minutes before no-news jumps. **(d)** Same plot for past binarised trends B_t, in linear scale. **(e)** Same plot for past trends T_t, in linear scale. Note that the averaging is performed on the absolute values of T_t and B_t.

- Binarized past price trend:

$$B_t = \kappa \frac{J_t}{|J_t|} + (1-\kappa) B_{t-1}\,,$$

using the same value of κ as above;

- Instantaneous average LOB sparsity s_t, defined using the 2 best limit prices:

$$s_t = \max\left[\frac{p_t^a - p_t^{b+1}}{\psi(1+\log V_t^b)}\,,\, \frac{p_t^{a+1} - p_t^b}{\psi(1+\log V_t^a)}\right]\,,$$

where ψ is the tick size (here $\psi = 0.01\$$ for all stocks in our sample). Note that we define the sparsity using the less dense side of the LOB. Given that s_t is sensitive to the local market activity and posses non-negligible intra-day periodicity, we define the associated z-score

$$S_t = \frac{s_t}{f_t \sigma_t} - \mu_t,$$

where σ_t is the standard deviation of s_t over a rolling window including the last day worth of data ($K = 390$), f_t is the average value of s_t/σ_t across

all the points with the same intra-day periodicity and μ_t is the average value of $s_t/f_t\sigma_t$ over a rolling window including the last day worth of data

In Fig. 3, we show how the average profiles of these five measures differs when calculated only on SEC clusters of jumps (not preceded by any news) or only on EMC clusters (in close proximity of a news). Averaging is done by shifting time such that for each cluster, $t = 0$ corresponds to the *first* jump of the cluster. Given that we are not interested in the sign of past trends, but only their magnitude, the averaging is performed on the absolute values of T_t and B_t.

The panels clearly show that SEC clusters are preceded by a slow increase of volatility and trends. The volatility increase starts to be statistically relevant up to 75 minutes before the occurrence of the first jump. The sparsity of the order book does increase, albeit weakly, 15 minutes before no-news jumps (see Fig. 3, panel c). We however expect that the final drop of liquidity takes place at higher frequency, due to the fierce competition between High-Frequency liquidity providers.

EMC clusters, on the other hand, happen much more abruptly, and their average profiles before the first jump hardly show any increase at all for all five metrics.

Consistent with observations for other social systems, we also see a clear difference in the relaxation of the volatility *after* the first jump. Endogenous clusters, even containing fewer jumps, revert to the average baseline volatility more slowly than exogenous clusters. Relaxation after EMC jumps is not only faster, it actually undershoots the baseline volatility: two hours after the first jump, four out of five indicators appear to be lower than the values recorded two hours before the first jump. This was also noted in Ref. [31], and interpreted by arguing that after the release of news, uncertainty about price is actually reduced. In contrast, endogenous jumps cannot be rationalized by market participants, and uncertainty remains high for a longer period.

Following previous work [3, 12, 15, 19, 20, 31] we now quantify the speed of the pre-jump and post-jump dynamical profiles by fitting a double power-law function of the form:

$$|J_t| = f(t) = \begin{cases} \frac{N_\ell}{|t-t_c|^{p_\ell}} + d, & (t < t_c) \\ \frac{N_r}{|t-t_c|^{p_r}} + d, & (t > t_c), \end{cases} \tag{3}$$

where d is the baseline volatility, $t_c \in [t_j - 1, t_j]$ is the time of the shock and t_j is the time at which the first jump of the cluster occurs. To estimate the coefficients appearing in Eq. (3), we use a non-linear least squares fitting and discard the first jump of a cluster. We then check that the normalized residuals are normally distributed using a Shapiro-Wilk test [55] at the 0.01 statistical significance.

For the exponents, we find $p_\ell = 0.36 \pm 0.02$ and $p_r = 0.40 \pm 0.02$ for SEC clusters and $p_\ell = 0.08 \pm 0.01$ and $p_r = 0.68 \pm 0.01$ for the EMC clusters. These values of p_r are different, but not very far, from those reported in [31], i.e. $p_r \approx 0.5$ for SEC and ≈ 1 for EMC.

For the amplitudes $N_{\ell/r}$, we find $N_\ell = 0.235 \pm 0.009$ and $N_r = 0.59 \pm 0.01$ for SEC clusters and $N_\ell = 0.68 \pm 0.01$ and $N_r = 4.76 \pm 0.06$ for EMC clusters. The estimated SEC baseline volatility is $d = 0.66 \pm 0.01$ while we find $d = 0.48 \pm 0.01$ for EMC. This discrepancy in the baseline volatility is because, as we can observe in Figure 3, the asymptotic post-cluster volatility of the EMC cluster is lower than the pre-shock baseline.[64] The estimated jump time of the SEC cluster $t_c = t_j \pm 0.08$ coincides with the time of the first jump of a cluster while, for the EMC class, we find $t_c = t_j - 1.00 \pm 0.03$, which is consistent with a pre-shock explosive growth.

Note that the relaxation of the LOB sparsity after a jump can also be fitted by a power-law with $p_r^S \approx 0.4$ for SEC jumps and $p_r^S \approx 0.7$ for EMC jumps, not very different from the ones governing the relaxation of $|J_t|$. On the other hand, the pre-jump SEC profiles start picking up too close to t_j to allow for a meaningful fit with our one minute resolution time.

C. Predictions of a Hawkes Model

Besides confirming the different volatility relaxation speeds between endogenous and exogenous jumps, as in other studied systems, our results show that there exist an asymmetry between the pre-jump growth and the post-jump relaxation *even for SEC*. To rationalize their findings Refs. [12, 13] postulate the existence of a self-exciting process of the form

$$\lambda(t) = \lambda_0(t) + \sum_{t_i < t} \phi(t - t_i) \ , \tag{4}$$

where $\lambda(t)$ the instantaneous rate of price moves, $\lambda_0(t)$ is the exogenous rate of price moves, t_i is the time at which previous price moves took place, and $\phi(\tau)$ is the memory kernel of the system, which captures the way past events enhance the probability of current events.

The rationale for using the self exiting process of Eq. (4) is fairly intuitive. Given the fact that the LOB is a public source of information, any change (exogenous or endogenous) in the instantaneous volatility may trigger a reaction in some market participants whose actions will have an effect on the volatility itself which will then trigger a second generation effect on other market participants and so on and so forth. This impact of a trader onto other traders, or of past volatility onto future volatility, is not instantaneous and it is modelled by the memory kernel $\phi(t - t_i)$.

By simply assuming a power-law memory function $\phi(\tau) \sim 1/\tau^{1+\theta}$ with $0 < \theta < 1$, Eq. (4) elegantly predicts two different profiles for exogenous and endogenous jumps [13, 56] when the process is marginally stable (i.e. when $n := \int_0^\infty \mathrm{d}\tau \phi(\tau) \to 1$). One finds the following

behaviour for the pre- and post-jump profiles:

$$|J_t| \propto \begin{cases} (t-t_{\rm j})^{\theta-1}, & \text{EMC}, \quad t > t_{\rm j}, t-t_{\rm j} \ll (1-n)^{-\frac{1}{\theta}}; \\ (t-t_{\rm j})^{-\theta-1}, & \text{EMC}, \quad t > t_{\rm j}, t-t_{\rm j} \gg (1-n)^{-\frac{1}{\theta}}; \\ |t-t_{\rm j}|^{2\theta-1}, & \text{SEC}, \quad t \lesseqgtr t_{\rm j}, \end{cases} \tag{5}$$

with a flat profile (no precursor) for EMC, $t < t_{\rm j}$.

Comparing these predicted profiles with the average ones plotted in Fig. 3, we see that a post-jump dynamics with $\theta = 0.3$ is consistent with our data for which $p_r^{SEC} \approx 0.4 = 1-2\theta$, and $p_r^{EMC} \approx 0.7 = 1-\theta$, positing, as argued in [33], that financial markets are indeed close to criticality ($n \approx 1$). Note that Wehrli et al. [57] have recently questioned this assumption, asserting that the low frequency kernel contribution to $n := \int_0^\infty {\rm d}\tau\phi(\tau)$ is dominated by the exogenous dynamics of the rate $\lambda_0(t)$ in Eq. (4). While this may well be the case, we satisfy ourselves in this work with the idea that critical Hawkes processes provide an *effective* description of feedback effects in financial markets, and treat Eq. (5) as a convenient fitting function.

Within this framework, the value we observe for θ is, quite remarkably, exactly the same as the one reported in previous studies on other social systems [12, 13]. Note that our value for the relaxation exponent $p_r^{SEC} \approx 0.4$ is quite close to the one estimated in Ref. [20], where post-jump volatility profiles of liquid US stocks were also studied (in the period 2000-2002).

However, the pre/post jump symmetry predicted by the model for SEC jumps is (mildly) violated – see the values of N_ℓ and N_r. We conjecture that such an asymmetry could be captured by the generalization of Eq. (4) recently proposed in [37], where not only past activity, but also past price trends, feedback on the current rate of activity. Such a coupling indeed leads to a measurable time reversal asymmetry in the volatility dynamics [37] and therefore can potentially produce an asymmetry between pre- and post-shock volatility dynamics as the one we observe [65]. We leave this question open for further investigation. It is worth mentioning that another possible explanation for such asymmetry is that a non negligible portion of the jumps we marked as endogenous are driven by exogenous information not detected by our news database.

IV. CLASSIFICATION OF SINGLE VOLATILITY PROFILES

In this final section we show that, even in a highly noisy environment such as financial markets, the classification of different jumps into SEC and EMC provided by the news feed can be successfully reconstructed only using individual volatility profiles.

Even if the Hawkes model (Eq. (4)) is not fully compatible with average profiles, as shown in the previous section, we attempt to use the functional forms suggested by Eq. (3) to fit *individual* volatility profiles and infer from such fits the nature of the observed events.

To do so, we fit the functional form of Equation 3 to each single volatility profile. In line with Refs. [12, 13] and with the discussion of the previous section, we expect the following characteristics:

- $p_r^{EMC} \approx 1-\theta > p_r^{SEC} \approx 1-2\theta$: higher relaxation exponents for those clusters which happen in close proximity of the release of a piece of news.

- $p_\ell^{EMC} \ll 1$ or $\gg 1$: the unanticipated, explosive nature of EMC jumps leads to a numerically determined exponent that is either very large or very small.

- Defining the asymmetry $\mathcal{A}$ of a jump from integrated area under the pre- and post- region of the profiles (see the Appendix for an operational definition) we expect $|\mathcal{A}^{EMC}| > |\mathcal{A}^{SEC}|$: the endogenous class is characterised by a rather symmetric pre-jump growth and post-jump relaxation, while exogenous events are strongly asymmetric.

Naturally, we do not expect such a sharp distinction between EMC and SEC at the level of individual events. For example, there are cases where the news leaks before announcement, leading to an increase of volatility ahead of the jump. Conversely, some endogenous events may show very little pre-jump activity since they can be triggered by "fat-fingers", by rogue algorithms or by some exogenous piece of information not present in our news database. Nevertheless, we will show that a relatively robust classification can still be performed by only considering the shapes of the single volatility profiles.

Given the high level of noise in price movements, we restrict our analysis to relatively high intensity clusters, i.e. to those clusters made up of at least two jumps. This leave us with a total of 10,491 clusters of jumps, out of which 391 happened in proximity of news releases. For each volatility profile, we perform a direct non-linear least squares fit of Eq. 3 (see the Appendix for a detailed description of the procedure) and we keep only those fits with a median relative error on the coefficients smaller than one. This leaves us with 5,461 SEC and 321 EMC events. In Figure 4, we plot the empirical distribution of the fitted values of p_ℓ, p_r and $\mathcal{A}$ for both types of events.

First of all, we observe that the results we obtain are remarkably consistent with the results on average profiles reported in the previous section. Indeed, one finds that the post-jump exponents p_r tend to be larger for EMC jumps than for SEC jumps; the difference $|p_\ell^{EMC} - 1|$ is large. Moreover, we see that, whereas the values of $\mathcal{A}^{SEC}$ are clustered around zero, the peak of the distribution of $\mathcal{A}^{EMC}$ is clearly shifted towards negative values, as expected. Moreover, we notice that the median values of the empirical p_r distribution are, respectively, 0.43 and 0.7 for the SEC and EMC jumps, values that are extremely close to the relaxation exponents found for the

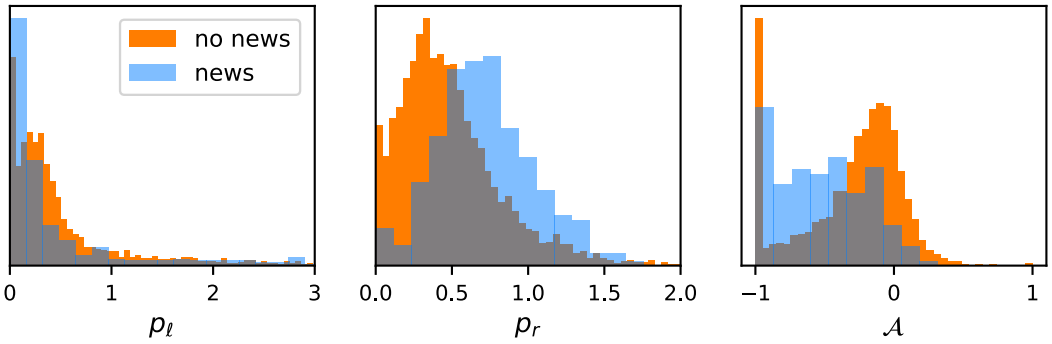

FIG. 4: Results of the double power-law fitting on the instantaneous volatility profiles. The name of the best fitting parameter each plots represents can be read directly on the plots.

	Logit	Probit
p_ℓ	-0.432***	-0.199***
	(0.080)	(0.036)
p_r	0.469***	0.300***
	(0.131)	(0.070)
$\mathcal{A}$	-1.897***	-0.906***
	(0.211)	(0.104)
const.	-3.623***	-2.001***
	(0.110)	(0.052)
AUC	0.73	0.73
pseudo-R^2_{adj}	0.069	0.070

Standard errors in parentheses. Two-tailed test.
*** $p < 0.001$

TABLE I: Results of the regression of p_ℓ, p_r and $\mathcal{A}$ against the class of each volatility time series obtained by using the news data. We use both a probit and a logit model to perform the regression.

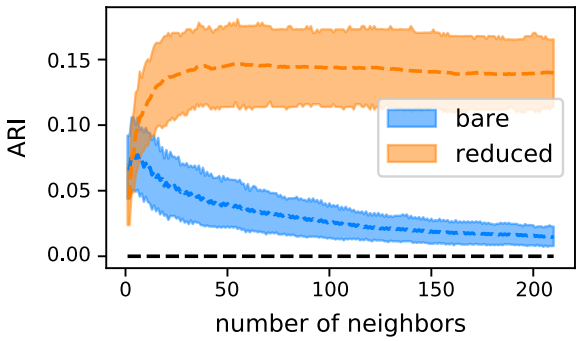

FIG. 5: Average ARI scores of a K-nn exploration of the space around each time series at different K levels. In blue we report the results when each time series is considered in its bare form, while in red we report the results obtained when each time series is embedded in the space $(p_\ell, p_r, \mathcal{A})$. The bands are the upper and lower 0.01 quantiles of the ARI distribution for a particular k when a bootstrapped subsample of the endogenous class is performed to match the number of elements in the exogenous class.

aggregated volatility profiles, and again consistent with the predictions of the Hawkes model 5 with $\theta = 0.3$.

In order to confirm that exogenous and endogenous events are genuinely characterised by different volatility profiles, we use the values of p_ℓ, p_r and $\mathcal{A}$ fitted on individual profiles to perform a regression on the given *a priori* EM and SE classes using both a Probit and a Logit model. The results of the two regressions are reported in Table I. First of all we observe that the classification task can successfully be performed using the three selected features, as witnessed by the values of the Area Under the Curve (AUC) and of the pseudo-R^2.

Moreover we see that the three features (higher pre-shock explosiveness, stronger asymmetry and faster post-shock relaxation) predicted for the EMC (marked as 1 in the regression) are indeed attested by the values of the parameters associated with each regressor.

In order to show that the results obtained are genuine, we randomly split our pool of jumps into a training and a test set (using a 80/20 ratio). We then train the regression models of Table I using only the former set and we test in on the latter (i.e. we test the model on unseen data) by means of the AUC metric. We repeat the process 1000 times. The average values of the out-of-sample AUC coming from this experiment are 0.72 ± 0.03 for both the Logit and Probit model, very close to the full sample result.

To further corroborate that the three features we have selected provide a meaningful low dimensional embedding for the classification of jumps into EM and SE classes, we explore the neighbourhood of each volatility profile by means of a distance weighted K-nearest neighbors algorithm [58]. First, we perform a bootstrap down-sampling of the SEC sample in order to have the same number of endogenous and exogenous time series. For each time series, we consider, using the euclidean distance, its first K neighbours and their (known) classes. We then assign each time series to the class most repre-

sented among those K neighbours. We repeat the process for each time series.

We then compare the resulting classification with the correct one. This comparison is performed using the so-called Adjusted Rand Index [59] (ARI), which is 0 for a random classifier. In Fig. 5 we plot the result of this K-nn exploration at various values of K. We observe that, for low K values, using the bare time series or its lower dimensional embedding gives comparable ARI values, which are low but both distinguishable from 0. As soon as we move away from $K = 1$ (and the so called "curse of dimensionality" kicks in for the full time series) we see that the average ARI of the K-nn classification becomes 0 for the full time series, while it rapidly converges to 0.15 for the three dimensional embedding we propose. This latter observation suggests that, in the space $(p_\ell, p_r, \mathcal{A})$, exogenous and endogenous jumps are overlapping but distinguishable clusters of points and therefore that SEC and EMC are distinguishable classes also when the intrinsic noise of the systems is not filtered out by an aggregation procedure.

To further strengthen our results, we have also performed the very same analyses using a fitting procedure, similar to the one suggested in Refs. [12, 13], and we find qualitatively similar results (see the Appendix for a detailed explanation).

V. CONCLUSIONS

Building upon the literature characterising the relaxation properties of financial systems after large exogenous shocks, we have argued that such fingerprints can fruitfully be used to disentangle exogenous and endogenous events, in close analogy with what has been observed in other social systems where self-exciting effects play an important role.

Using 5 years of minute by minute data collected from the Limit Order Books of 300 different NYSE stocks, we have shown that the average characteristics of clusters of abnormally large price variations in close proximity of a news release differ significantly from those occurring without any triggering event in the news feed. In particular, we have shown how the *average* profiles of the instantaneous volatility, past volatility and normalised past trends all display specific fingerprints that discriminate between exogenous and endogenous jumps. We have also focused on *individual* volatility profiles and have shown that, despite a modest signal to noise ratio, the parameters of the fitted power-laws allow one to reconstruct the classification provided by the news feed of large jumps into the "self-excited" and news induced class.

Whereas the existence of exogenous and endogenous types of shocks in financial systems may appear natural to many, it should be stressed that it is still a matter of intense skepticism in the current economic literature, given the difficulty to reconcile this view with the enduring Efficient Market Theory. We hope that the present work will help convince researchers that, while markets do indeed strongly and rapidly react to outstanding news, small and seemingly unimportant fluctuations may lead to a cascade of events that trigger large price jumps. In fact, most jumps appear to be of such type – market participants do endemically interact, both directly and indirectly. The very existence of public sources of information – such as the price itself and the Limit Order Book – leads to global interactions and destabilising feedback loops.

Our study can be seen as supporting a micro-structural interpretation of the excess volatility puzzle [60]: if large price jumps can appear out of the blue, as a result of intrinsic market fragility, then it is not surprising that prices are also too volatile. In fact, dissecting the mechanisms that lead to clusters of jumps using a multi-dimensional Quadratic Hawkes processes calibrated on tick-by-tick order data would be a very interesting follow up which we leave for future work (see [61]).

Finally, while we are confident that our results have a large degree of universality, extending them to other asset classes, market places or timescales would certainly be of interest and would bolster our claim that endogenous price jumps fall in the wider class of self-excited events.

Acknowledgments

We thank Pierre-Philippe Crépin who contributed to the early stages of this work, as well as Cecilia Aubrun, Charles-Albert Lehalle, Antoine Fosset and Iacopo Mastromatteo for fruitful discussions. We also thank Gary Kazantsev (Bloomberg), Fabrizio Lillo and Didier Sornette for encouragements, comments and suggestions.

This research was conducted within the Econophysics & Complex Systems Research Chair, under the aegis of the Fondation du Risque, the Fondation de l'Ecole polytechnique, the Ecole polytechnique and Capital Fund Management.

[1] P. Bak, *How nature works: the science of self-organized criticality* (Springer Science & Business Media, 2013).
[2] S. Albeverio, V. Jentsch, and H. Kantz, *Extreme events in nature and society* (Springer Science & Business Media, 2006).
[3] D. Sornette, *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools* (Springer Science & Business Media, 2006).
[4] N. N. Taleb, *The black swan: The impact of the highly improbable*, vol. 2 (Random house, 2007).
[5] D. Sornette, in *Extreme events in nature and society* (Springer, 2006), pp. 95–119.

[6] D. Sornette and A. Helmstetter, Physica A: Statistical Mechanics and its Applications **318**, 577 (2003).
[7] A. G. Hawkes, Biometrika **58**, 83 (1971).
[8] A. Helmstetter and D. Sornette, Geophysical Research Letters **30** (2003).
[9] E. Bacry, I. Mastromatteo, and J.-F. Muzy, Market Microstructure and Liquidity **1**, 1550005 (2015).
[10] G. Mohler et al., The Annals of Applied Statistics **7**, 1525 (2013).
[11] D. Lando and M. S. Nielsen, Journal of Financial Intermediation **19**, 355 (2010).
[12] R. Crane and D. Sornette, Proceedings of the National Academy of Sciences **105**, 15649 (2008).
[13] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, Physical Review Letters **93**, 228701 (2004).
[14] F. Deschâtres and D. Sornette, Phys. Rev. E **72**, 016112 (2005).
[15] F. Lillo and R. N. Mantegna, Physical Review E **68**, 016119 (2003).
[16] A. M. Petersen, F. Wang, S. Havlin, and H. E. Stanley, Phys. Rev. E **82**, 036114 (2010), URL `https://link.aps.org/doi/10.1103/PhysRevE.82.036114`.
[17] P. Weber, F. Wang, I. Vodenska-Chitkushev, S. Havlin, and H. E. Stanley, Phys. Rev. E **76**, 016109 (2007), URL `https://link.aps.org/doi/10.1103/PhysRevE.76.016109`.
[18] F. Lillo and R. N. Mantegna, Physica A: Statistical Mechanics and its Applications **338**, 125 (2004).
[19] A. M. Petersen, F. Wang, S. Havlin, and H. E. Stanley, Physical Review E **81**, 066121 (2010).
[20] Á. G. Zawadowski, G. Andor, and J. Kertész, Quantitative Finance **6**, 283 (2006).
[21] T. Utsu, Geophys. Mag. **30**, 521 (1961).
[22] S.-H. Poon and C. W. Granger, Journal of economic literature **41**, 478 (2003).
[23] K. Yamasaki, L. Muchnik, S. Havlin, A. Bunde, and H. E. Stanley, Proceedings of the National Academy of Sciences **102**, 9424 (2005).
[24] X. Jiang, T. Chen, and B. Zheng, Physica A: Statistical Mechanics and its Applications **392**, 5369 (2013).
[25] A. Ponzi, F. Lillo, and R. N. Mantegna, Phys. Rev. E **80**, 016112 (2009), URL `https://link.aps.org/doi/10.1103/PhysRevE.80.016112`.
[26] R. Hisano, D. Sornette, T. Mizuno, T. Ohnishi, and T. Watanabe, PloS one **8**, e64846 (2013).
[27] M. Rambaldi, P. Pennesi, and F. Lillo, Phys. Rev. E **91**, 012819 (2015).
[28] D. M. Cutler, J. M. Poterba, and L. H. Summers, Tech. Rep., National Bureau of Economic Research (1988).
[29] R. C. Fair, The Journal of Business **75**, 713 (2002).
[30] C. Hopman, Quantitative Finance **7**, 37 (2007).
[31] A. Joulin, A. Lefevre, D. Grunberg, and J.-P. Bouchaud, Wilmott Magazine **46** (2008).
[32] V. Filimonov and D. Sornette, Physical Review E **85**, 056108 (2012).
[33] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud, The European Physical Journal B **86**, 1 (2013).
[34] S. Wheatley, A. Wehrli, and D. Sornette, Quantitative Finance **19**, 1165 (2019).
[35] S. Koyama and S. Shinomoto, Physical Review Research **2**, 043358 (2020).
[36] D. Sornette, Y. Malevergne, and J.-F. Muzy, in *The Application of Econophysics* (Springer, 2004), pp. 91–102.
[37] P. Blanc, J. Donier, and J.-P. Bouchaud, Quantitative Finance **17**, 171 (2017).
[38] A. Fosset, J.-P. Bouchaud, and M. Benzaquen, Journal of Statistical Mechanics: Theory and Experiment **2020**, 063401 (2020).
[39] S. Stoikov, Quantitative Finance **18**, 1959 (2018).
[40] A. Tripathi, V. Null, and A. Dixit, Qualitative Research in Financial Markets **12**, 505 (2020).
[41] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison, Quantitative Finance **13**, 1709 (2013).
[42] J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, *Trades, quotes and prices: financial markets under the microscope* (Cambridge University Press, 2018).
[43] G. Bormetti, L. M. Calcagnile, M. Treccani, F. Corsi, S. Marmi, and F. Lillo, Quantitative Finance **15**, 1137 (2015).
[44] W. Horbelt and J. Timmer, Physics Letters A **310**, 269 (2003).
[45] S. S. Lee and P. A. Mykland, The Review of Financial Studies **21**, 2535 (2008). [45][46] Jump-robust estimator of the local volatility
[46] K. Boudt, C. Croux, and S. Laurent, Journal of Empirical Finance **18**, 353 (2011). [46][47] Non-param model to determine which change in mid-price is considered abnormal.
[47] R. Cont, Quantitative Finance **1**, 223 (2001).
[48] E. Bacry, J. Delour, and J.-F. Muzy, Physical Review E **64**, 026103 (2001).
[49] A. Chronopoulou and F. G. Viens, Quantitative Finance **12**, 635 (2012).
[50] K.-I. Goh and A.-L. Barabási, EPL (Europhysics Letters) **81**, 48002 (2008).
[51] A. Saichev and D. Sornette, Physical Review Letters **97**, 078501 (2006).
[52] R. Chicheportiche and A. Chakraborti, Physica A: Statistical Mechanics and its Applications **474**, 312 (2017).
[53] K. Boudt and M. Petitjean, Journal of Financial Markets **17**, 121 (2014). Remove any cluster of jumps happening within 100 minutes to account for contamination effect
[54] M. G. Kendall, Biometrika **30**, 81 (1938).
[55] S. S. Shapiro and M. B. Wilk, Biometrika **52**, 591 (1965).
[56] A. Helmstetter and D. Sornette, Journal of Geophysical Research: Solid Earth **107**, ESE (2002).
[57] A. Wehrli, S. Wheatley, and D. Sornette, Quantitative Finance **21**, 729 (2021), https://doi.org/10.1080/14697688.2020.1838602, URL `https://doi.org/10.1080/14697688.2020.1838602`.
[58] N. S. Altman, The American Statistician **46**, 175 (1992).
[59] N. X. Vinh, J. Epps, and J. Bailey, The Journal of Machine Learning Research **11**, 2837 (2010).
[60] R. J. Shiller, The American Economic Review **71**, 421 (1981).
[61] A. Fosset, J.-P. Bouchaud, and M. Benzaquen, The European Journal of Finance **0**, 1 (2021), https://doi.org/10.1080/1351847X.2021.1917441, URL `https://doi.org/10.1080/1351847X.2021.1917441`.
[62] L. M. Calcagnile, G. Bormetti, M. Treccani, S. Marmi, and F. Lillo, Quantitative Finance **18**, 237 (2018). [62] rescale returns after missing period
[63] J. Beran, *Statistics for long-memory processes*, vol. 61 (CRC press, 1994).
[64] Leaving d_r and d_ℓ free does not significantly change the values of the exponents p_r and p_ℓ.
[65] Note that, as measured in Refs. [37, 38] the strength of the feedback between past trends and future volatility is much smaller than the one between past and future volatility. This is compatible with the observed weak degree of asymmetry between pre- and post-jump profiles.

[37][38] Exponential moving average of past excess volatility.

Appendix A: Data handling and descriptive statistics

1. Financial data preprocessing

Why did they select these stocks? They mention something about the turnover from 2019.

The full list of the stocks' tickers included in our analysis is the following: TSLA, AMZN, AAPL, MSFT, FB, NVDA, GOOGL, GOOGL, NFLX, AMD, ZM, BA, INTC, V, PYPL, ADBE, JPM, BRK/A, MA, BAC, CSCO, JNJ, DIS, UNH, QCOM, CMCSA, COST, CRM, PG, XOM, GILD, MU, T, PEP, PFE, HD, ROKU, C, AVGO, BKNG, WMT, TXN, MRNA, AMGN, CHTR, SBUX, WFC, VZ, MRK, UAL, CVX, KO, BYND, LRCX, MCD, REGN, TMUS, DOCU, ABBV, ORCL, SQ, NKE, BMY, BIIB, ATVI, PTON, AMAT, EBAY, AAL, TMO, IBM, FISV, VRTX, GS, NEE, ISRG, UBER, INTU, UTX, NOW, LLY, LULU, CRWD, UNP, ABT, TTD, HON, LMT, ILMN, MMM, MS, MELI, TWTR, LOW, AMT, MDLZ, TGT, ADP, DXCM, ADI, EQIX, GE, BLK, EA, WDAY, ADSK, DAL, MAR, CME, UPS, XLNX, CSX, DHR, CAT, SPGI, SPLK, FIS, TWLO, PM, CVS, NEM, AXP, COUP, WYNN, ORLY, TDOC, WBA, FDX, BDX, SNAP, ECL, NVAX, TJX, ETSY, PLD, F, EXPE, MCHP, SCHW, ROST, OKTA, KLAC, SWKS, ANTM, CI, DKNG, DDOG, DG, GM, MO, SPG, DLR, CMG, NKLA, ALGN, DUK, CTXS, XEL, EL, TTWO, CCI, CL, COP, OXY, HUM, EXC, CTSH, SHW, VIAC, NOC, ALXN, WORK, BSX, APD, ULTA, ZS, WDC, ENPH, MCK, D, SBAC, LUV, GPN, LYFT, ZTS, CLX, PENN, ICE, DE, W, SYK, USB, DD, KMB, MXIM, SO, DLTR, KR, DPZ, PINS, MPC, MTCH, KHC, CDNS, MNST, SNPS, AEP, LHX, INO, ZG, ITW, SEDG, NSC, PNG, TROW, FTNT, FSLY, IAC, IDXX, MET, EW, RNG, AKAM, CNC, CTAS, TIF, VLO, GD, FAST, PANW, LVGO, MMC, EOG, PAYX, WM, HCA, AZO, PSX, MCO, PGR, QRVO, TER, CSGP, FCX, TSCO, MDB, PCAR, GIS, TFC, LVS, ANSS, BAX, PSA, BK, CZR, VEEV, HPQ, VRSN, SRE, HLT, TDG, CMI, CERN, ROP, VRSK, STZ, EMR, MGM, KMI, SGEN, CPRT, TRV, DHI, DOW, CVNA, COF, SYY, PLUG, INCY, MSI, ALL, CHRW, AIG, FLT, AVB, FE, MSCI, PPG, BBY, RUN, BMRN, YUM, WMB, QDEL, ODFL, ED, LEN, PXD, PAYC, NLOK, AMTD.

For each stock, we have a complete description of the first 6 price levels of its LOB. If we record a missing value in the volumes or the prices at any price level which is not the best bid/ask, we simply consider the associated LOB sparsity as missing and we exclude it from any calculation performed in the main text. If we record a missing value at the best bid or at the best ask, we move the last available observation forward in time. If, in doing so, we obtain an impossible price level (i.e. the price at the best bid/ask is lower/greater than the price at the second best or greater/lower than the price at the best ask/bid), we mark as missing the associated mid-price. We also mark as missing any mid-price associated with a minute when both the best bid and best ask prices are missing. We exclude from our analysis any day with more than 25 consecutive missing mid-prices. We also exclude from our analysis the days with more than 25 consecutive minutes without any recorded price movement and the days that do not have at least 300 minutes with a recorded price movement (i.e. a return which is not 0 nor missing). To further discount the possibility of detecting a spurious jumps due to an interval of missing values, we follow Ref. [62] and we rescale the returns computed after a missing period with the square root of the period length. For example, given the price series $p_0, p_1, -, -, p_4, p_5$, we construct the following log-returns series $\log\frac{p_1}{p_0}, \mathrm{NA}, \mathrm{NA}, \frac{1}{\sqrt{3}}\log\frac{p_4}{p_1}, \log\frac{p_5}{p_4}$.

2. Returns standardization

As mentioned in the main text, the log-returns r_t time series in not suitable for the identification of price jumps and must be standardize by passing from r_t to $J_t = \frac{r_t}{\sigma_t f_t}$. As a jump-robust estimator of the local volatility, we use, as suggested in Refs. [45, 46], the square root of the average realised bipower variation:

$$\sigma_t^2 = \frac{\pi}{2K}\sum_{i=1}^{K} |r_{t-i}|\,|r_{t-i+1}| \,.$$

To estimate the periodicity component f_t of the volatility, we perform a two step procedure based on Refs. [45, 46]. First of all, let's define $\hat{r}_t = \frac{r_t}{\sigma_t}$. Let $\hat{r}_{1,i}, \ldots, \hat{r}_{n_i,i}$ be the set of standardized returns having the same periodicity factor as r_i, i.e. the returns of a given stock, all recorded at the given time interval i. We define the periodicity factor f_i of the time interval i as:

$$f_i = \frac{W_i}{\sqrt{T^{-1}\sum_j W_{i-j}^2}}, \quad \text{where} \quad W_i = \sqrt{1.081\frac{\sum_j^{n_i}\Theta(-\hat{r}_{j,i}^2+x)\hat{r}_{j,i}^2}{\sum_j^{n_i}\Theta(-\hat{r}_{j,i}^2+x)}} \,.$$

Beside being normalised so that the squared periodicity factor has mean one over any local window of length T, f_i is a simple weighted standard deviation of the squared standardized returns $\hat{r}_{j,i}$. The weights are 1 for those squared

standardize returns which are below a threshold value x and 0 otherwise. To estimate the periodicity we perform a two-fold procedure. First we estimate f_i^0 using $x = 4^2$, i.e. by excluding from the calculation those rescaled returns $\hat{r}_t$ more than 4 standard deviation away from the average. Then, using $\hat{r}_t/f_i$ (which are now very close to be normally distributed), we perform a second periodicity estimation f_i^1 with a threshold value $x = 6.635$, i.e. the 99% quantile of the χ^2 distribution with one degree of freedom. We then define the final periodicity factor f_t as $f_t = f_t^1 f_t^0$. We use as periodicity cycle one day and we therefore consider as having the same periodicity factor all the returns of a stock which happen in the same minute of different days. In Panel **(c)** of Figure 6 we plot the estimated periodicity factors of two different stocks. As it can be seen, the two-fold procedure we used puts a final higher factor on the first 15 minutes of the day and leaves the rest almost untouched. It is therefore less prone to detect jumps at the opening of the trading day. We also remind that, to further discount for spourious effects due to the opening and closing, we excluded from the jump pool those jumps detected in the first/last 15 minutes of the day. We also explored the possibility of having a cycle of one week (i.e. we consider as having the same periodicity factor all the returns of a stock which happen in the same minute of the same day of different weeks) but we do not find any sizable difference in the final J_t.

To show that the final jump statistics J_t is indeed effective, in Panel **(a)** of Figure 6, we plot the autocorrelation function of $|J_t|$ for two different stocks. As it can be seen, J_t does a fairly good job in taking out from the volatility most of its seasonal components as well as most of its internal dynamics. To give the reader a better understanding of the jump statistics we use, in Panel **(d)**, we plot the final probability density function of $|J_t|$ with respect to that of the absolute standard normal distribution. Finally, in Panel **(b)** we show how the unconditional jump probability P_J resulting from J_t evolves over time for two selected stocks.

3. News data

In Panel **(a)** of Figure 7 we display the number of stock-specific news for each minute of the day. In Panel **(b)** and **(c)**, we respectively report the empirical distribution of news per stock and the average number of news per stock.

Appendix B: Power law fitting

In this section we detail the fitting procedure used in the main text as well as the secondary fitting procedure. First of all it should be noticed that fitting power law scaling laws is a subtle topic that has been vastly debated in the literature [63]. Here, we are trying to fit a double power law function to very noisy data, as such every step should be done with extreme care.

The secondary procedure we adopt is similar to the one suggested in Refs. [12, 13]. We consider the following functional form:

$$f_1(t|N_\ell, N_r, p_\ell, p_r, t_c) = \Theta(-t+t_c)\log\frac{N_\ell}{|t-t_c|^{p_\ell}} + \Theta(t-t_c)\log\frac{N_r}{|t-t_c|^{p_r}} \tag{B1}$$

Note that no baseline volatility d is added to the fit. This is done in order to keep the problem analytically solvable. To amend for this lack of a constant parameter, instead of fitting Eq. B1 to $\log|J_t|$, we fit it against $\log\frac{|J_t|}{J_0}$, i.e. against the instantaneous volatility profile normalised for the size of the first jump of a cluster. Doing so minimizes the influence of the baseline volatility on the fit, normalises all the fits and does not modify the values of the best fitting exponents of the power laws. Performing a least square fit of Eq. B1 on an empirical volatility series $\log\frac{|J_{t_i}|}{J_0}$ means to solve the following optimization problem:

$$\min_{N_\ell, N_r, p_\ell, p_r, t_c} \sum_i \left(f_1(t_i) - \frac{|J_{t_i}|}{J_0} \right)^2 .$$

Calling $|t_i - t_c| = \Delta t_i$, $A_{\ell/r} = \log N_{\ell/r}$, $L = \sum_{i|t_i<t_c} 1$, $R = \sum_{i|t_i>t_c} 1$, $s_\ell = \sum_{i|t_i<t_c} \log\Delta t_i$, $s_r = \sum_{i|t_i>t_c} \log\Delta t_i$, $S_\ell = \sum_{i|t_i<t_c} \log^2\Delta t_i$, $S_r = \sum_{i|t_i>t_c} \log^2\Delta t_i$, $D_\ell = \sum_{i|t_i<t_c} \frac{|J_{t_i}|}{J_0}$, $D_r = \sum_{i|t_i<t_c} \frac{|J_{t_i}|}{J_0}$, setting the partial derivatives with respect to $A_{\ell/r}, p_{\ell/r}$ to zero and solving the system of equations, gives:

$$p_\ell^\star = D_\ell \frac{L - s_\ell}{s_\ell^2 - L S_\ell} \ , \ A_\ell^\star = D_\ell \frac{s_\ell - S_\ell}{s_\ell^2 - L S_\ell} \ , \ p_r^\star = D_r \frac{R - s_r}{s_r^2 - R S_r} \ , \ A_r^\star = D_r \frac{s_r - S_r}{s_r^2 - R S_r} \ .$$

Note that these values of the best-fitting parameters are all functions of the unknown $t_c^\star$. To find the best-fitting shock time $t_c^\star$ able to minimize the sum of the squared residuals, we perform a numerical grid search inside the open

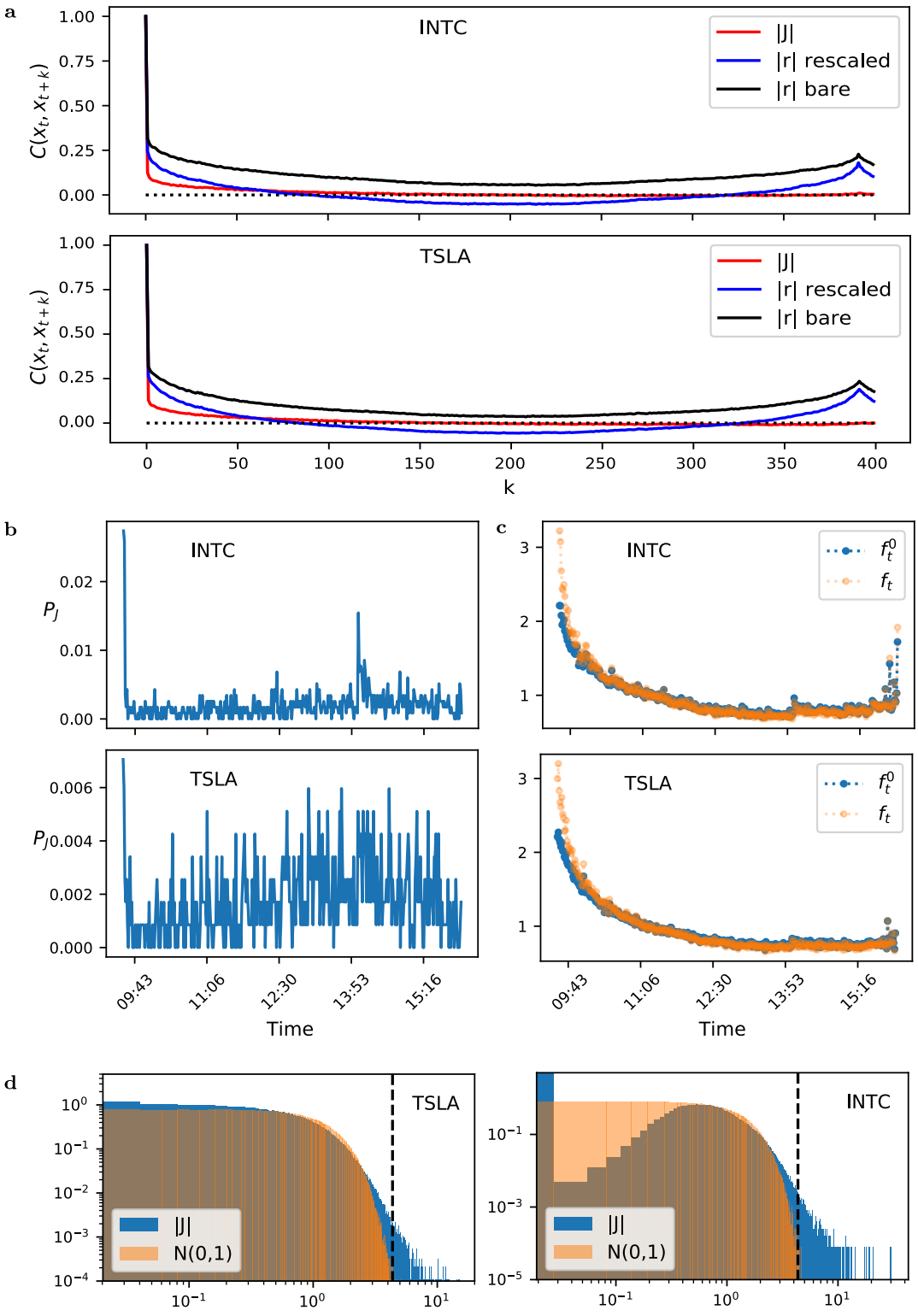

FIG. 6: **(a)** Autocorrelation functions of the absolute returns $|r_t|$ (black), of the absolute rescaled returns $|r_t|/\sigma_t$ (blue) and of the absolute jump statics $|J_t|$ (red) at different lags for TSLA and INTC stocks. **(b)** Evolution of the unconditional jump probability for TSLA and INTC stocks. **(c)** Estimated periodicity factors f_t for TSLA and INTC stocks. **(d)** Probability distribution of the absolute value of the jump statics J for two selected stocks. In orange we display the pdf of the absolute value of a standard normal distribution.

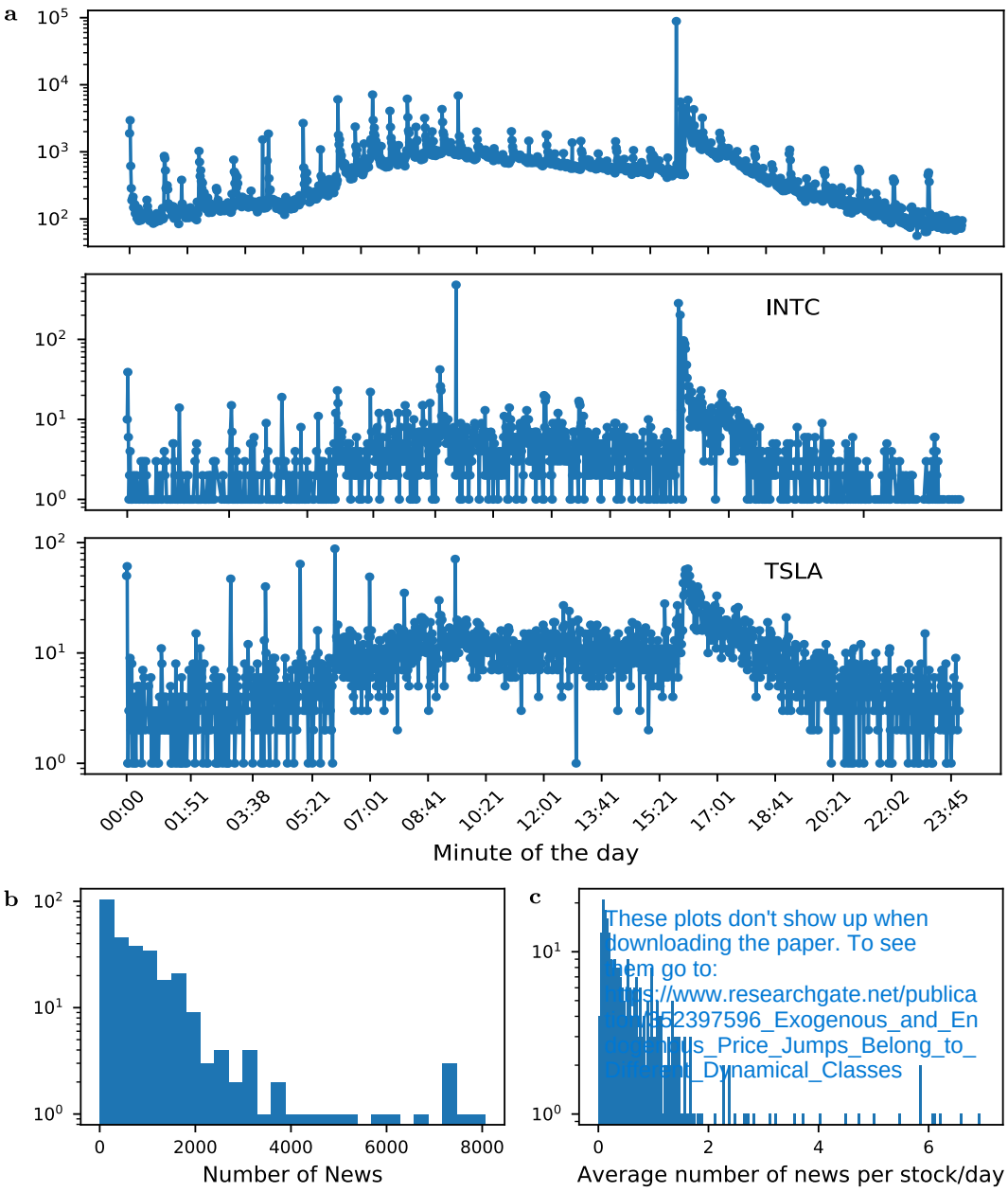

FIG. 7: **(a)** Number of news recorded for each minute of the day across all stocks and for two selected stocks. **(b)** Empirical distribution of the number of news (recorded within each trading day) across stocks. **(c)** Empirical distribution of the average daily number of news across stocks.

interval $(t_j - 1, t_j)$, where with t_j we indicate the time of the first jump of a cluster of jumps. The lest square fit defined in this way is unique given an empirical time series J_{t_i} which is also uniquely defined by an interval $[t_1, t_N]$. As such, we fix the fitting interval to a centered interval of length 160 minutes around t_j.

After finding the optimal values of p_ℓ, p_r and $\mathcal{A}$, we perform with them both the logistic regression exercise and the k-nn exploration detailed in the main text. As it can be seen from Table and Figure, the results we obtain are consistent with the one reported in the main text.

The procedure adopted in the main text to find the best fitting parameters is a non-linear lest square performed using the SciPy Python package. Fitting directly the functional form:

$$f_2(t|N_\ell, N_r, p_\ell, p_r, t_c, d) = \frac{N_\ell\ \Theta(-t+t_c)}{|t-t_c|^{p_\ell}} + \frac{N_r\ \Theta(t-t_c)}{|t-t_c|^{p_r}} + d\ ,$$

	Logit	Probit
p_ℓ	-15.06***	-7.7379***
	(2.13)	(1.04)
p_r	21.53***	10.86***
	(1.99)	(0.97)
$\mathcal{A}$	-16.91***	-8.24***
	(1.73)	(0.82)
const.	-7.87***	-3.97***
	(0.25)	(0.12)
AUC	0.82	0.82
pseudo-R^2_{adj}	0.198	0.198

Standard errors in parentheses. Two-tailed test.
*** $p < 0.001$

TABLE II: Results of the regression of $p^\star_\ell$, $p^\star_r$ and $\mathcal{A}$ against the class of each volatility time series obtained by using the news data. We use both a probit and a logit model to perform the regression.

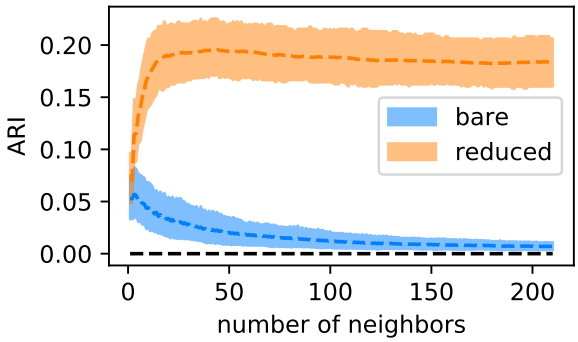

FIG. 8: Average ARI scores of a K-nn exploration of the space around each time series at different K levels. In blue we report the results when each time series is considered in its bare form, while in red we report the results obtained when each time series is embedded in the space $(p_\ell, p_r, |\mathcal{A}|)$. The bands are the upper and lower 0.01 quantiles of the ARI distribution for a particular k when a bootstrapped subsample of the endogenous class is performed to match the number of elements in the exogenous class.

on the empirical volatility series $|J_{t_i}|$ can become hard given the notoriously low signal-to-noise ratio of financial data. As such, we fit its cumulative sum $F_2(t_i) = \sum_{k=1}^{i} f_2(t_k)$ to $D_i = \sum_{k=1}^{i} |J_{t_k}|$. We restrict $t_c \in (t_\mathrm{j} - 1, t_\mathrm{j})$ and $d > 0$.

Appendix C: Market-wide jump detection

In order to mark a cluster of jumps as market-wide, one possibility (not used in the main text) is to compare the empirical number of clusters (of other stocks) it overlaps, against the one expected under a null-hypothesis of cluster independence. In order to create a null model of independent clusters, we perform the following randomization procedure. We call the beginning of a cluster, the time t_l at which the first jump of a cluster is recorded. We call the ending of a cluster, the time t_L at which the last jump of a cluster is recorded. Finally, we refer to $L = t_L - t_\mathrm{l}$ as the length of a cluster. Given a cluster of jumps C, we fix its position $[t^C_\mathrm{l}, t^C_L]$ and we perform a numerical shuffling of all the remaining clusters beginning and ending positions so that no cluster of the same stock can overlap, the length of each cluster is preserved and no cluster may be moved outside the month/year it has been observed. Once this is done, we compare at the 0.05 statistical significance the empirical number of cluster overlapping with C against the ones expected under our null model. Whenever we observe a cluster C with an higher number of overlaps then those accounted for by our null model, we mark it as market-wide.

Appendix D: Definition of $\mathcal{A}$

As a time asymmetry measure $\mathcal{A}$ we consider the following formula:

$$\mathcal{A} = \frac{A_\ell - A_r}{A_\ell + A_r} \ , \tag{D1}$$

where $f^*(t)$ is the best fitting curve 3, $A_\ell = \sum_{t=t_{\min}}^{t_j - 1} f^*(t)$ and $A_r = \sum_{t=t_j}^{t_{\max}} f^*(t)$. Equation D1 simply compares the area of the fitted curve before the shock time t_c^* with the normalised area after the shock.