Filtering the Intensity of Public Concern from Social Media Count Data with Jumps

Matteo Iacopini, Carlo Romano Marcello Alessandro Santagiustina

Received: 24 December 2019 | Accepted: 28 March 2021

DOI: 10.1111/rssa.12704

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Filtering the intensity of public concern from social media count data with jumps

Matteo Iacopini[1,2] | **Carlo R.M.A. Santagiustina**[3,4]

[1]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

[2]Tinbergen Institute, Amsterdam, The Netherlands

[3]Ca' Foscari University of Venice, Venice, Italy

[4]Venice International University, Venice, Italy

Correspondence
Matteo Iacopini, The Netherlands Tinbergen Institute, Vrije Universiteit Amsterdam, Amsterdam, Noord-Holland, The Netherlands.
Email: m.iacopini@vu.nl

Funding information
H2020 Future and Emerging Technologies, Grant/Award Number: 732942; H2020 Marie Sklodowska-Curie Actions, Grant/Award Number: 887220

Abstract

Count time series obtained from online social media data, such as Twitter, have drawn increasing interest among academics and market analysts over the past decade. Transforming Web activity records into counts yields time series with peculiar features, including the coexistence of smooth paths and sudden jumps, as well as cross-sectional and temporal dependence. Using Twitter posts about country risks for the United Kingdom and the United States, this paper proposes an innovative state space model for multivariate count data with jumps. We use the proposed model to assess the impact of public concerns in these countries on market systems. To do so, public concerns inferred from Twitter data are unpacked into country-specific persistent terms, risk social amplification events and co-movements of the country series. The identified components are then used to investigate the existence and magnitude of country-risk spillovers and social amplification effects on the volatility of financial markets.

KEYWORDS

Bayesian inference, count time series, jumps, online social media, particle filtering, risk perception

1 | INTRODUCTION

With globalization, the need to understand the dynamics of country-risk perception and its effects on financial markets has become a relevant issue to investors, central bankers and governments for both portfolio diversification and debt issuance (Campbell et al., 2001; Hassan et al., 2003; Huber et al., 2019). Major political events are known to be an important driver of market volatility (Bialkowski et al., 2008); however, a measure of the underlying country-risk perception is still lacking.

The effect of new information on volatility in international markets has been studied extensively in the past few years. Previously, it was held that global risks are the only considerable risks in financial markets, but more recently, researchers have begun to examine country risks and global risks to uncover local factors that cause stock market return volatility. Moreover, options have been traded on the VIX index since 2006, thus allowing volatility to be considered an asset.

The identification of public concerns related to specific issues of interest (e.g. country risks) requires a data source that can be used to proxy the perception of the general public. Unfortunately, existing sources of information on country-specific risks are either qualitative (World Economic Forum, 2017) or low frequency (Howell, 2011) and often represent the perceptions of small panels of experts, thus ruling out the possibility of using them as proxies for the general public's concerns.

To address this gap, in this paper, we focus on count series generated from big textual data regarding country-specific risks via online social media. This type of data is characterized by sudden information avalanches and is affected by topic-specific trends, thus requiring adequate statistical tools to disentangle these components. In this paper, we propose a new state space model for Web count data that allows country-specific risk series to be categorized into three separate components, each capturing a specific feature driving the intensity of the observables. In particular, (i) a smooth term represents the latent intensity of country-specific risk concerns, while (ii) a multiplicative jump component captures their social amplification, and (iii) a global component controls for common factors. This method of decomposing country-risk perception offers new ways to evaluate the impact of social media phenomena on market and sociopolitical systems. In this paper, we focus on the analysis of risk perception issues, but the proposed framework enables the use of count time series to analyse any Web phenomenon characterized by social amplification.

To overcome the limitations of the aforementioned data sources, we investigate user-generated Web data. Content generated by the Web community, such as online social media post-writing and sharing activities, have been widely used to track aggregate behaviour and infer the dynamics of the public's attention, perceptions and concerns about specific topics of interest (Rogers, 2013). Compared to other data sources (e.g. surveys and official statistics), big textual data from the Web allow for (i) larger sample sizes, (ii) higher velocity and (iii) real-time collection (Varian, 2014).

Data from social media generally consist of a collection of character strings, which can be transformed into other data types for analysis with standard statistical methods (Einav & Levin, 2014). For example, count data can be generated from texts through filtering conditions, which may consist of regular expressions (RegEx) or metadata restrictions (e.g. the author, geolocation or date). Count data from the Web are becoming increasingly valuable and widely used in several fields. For example, data from Google's search engine and Twitter have been used recently to improve forecasting accuracy in macroeconomics (D'Amuri & Marcucci, 2017) and finance (Ranco et al., 2015). The analysis of Web-based count data requires suitable and simple statistical tools designed for analysis and making predictions.

Figure 1 shows the daily number of Twitter posts published from April 2016 to March 2017 that contain the term '*risk*' and refer to the United States (US) or to the United Kingdom (UK). The series share some peculiar features: (i) smooth evolution and (ii) several jump events. The observed

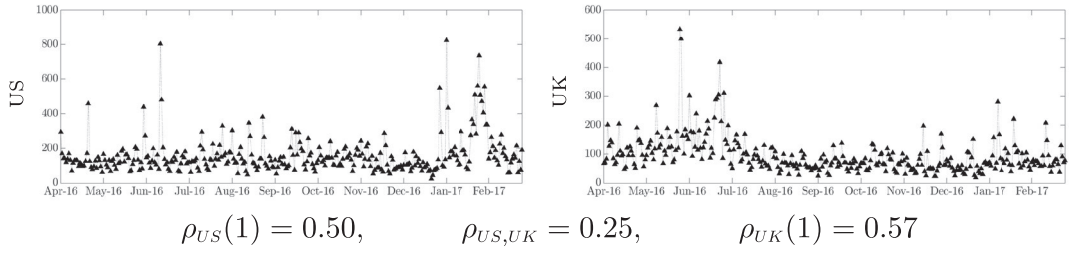

FIGURE 1 Top: daily time series of counts of Twitter posts from 1 April 2016 to 1 March 2017 for the US (left) and the UK (right). Bottom: autocorrelation of order 1 and instantaneous cross-correlation between the US and UK

series are also characterized by positive auto- and cross-correlation, thus suggesting the presence of both temporal and cross-sectional dependence. Social media users' concerns regarding global-scale risks (e.g. trade wars, global financial crisis and pandemics) may also impact country-specific activity, thus inducing the positive cross-correlation between the observed series in Figure 1. To account for this fact, we enlarge our sample by including a third series regarding global risks, which are not imputable to a specific country, and use it as a common driver of US and UK observables. State space (or parameter-driven) models provide a flexible and interpretable framework for the structural analysis of serially dependent time series (Durbin & Koopman, 2012). To account for the correlation structure of the data, we propose a new state space model in which the smooth temporal evolution of the observed variables is driven by a dynamic latent series representing social media users' concerns.

A parameter-driven Bayesian model for multivariate count data was recently proposed by Aktekin et al. (2018). In their framework, the cross-sectional sum of past observations affects the dynamics of all latent series. Similarly, Zaman et al. (2014) proposed a method for predicting the time path of retweets. However, our goal is to decompose the public' latent concerns regarding country-specific risks.

An additional distinctive feature of the data in Figure 1 is the presence of country-specific and non-synchronized activity peaks. These sudden jumps cluster over time and have different degrees of persistence and scale. Time series characterized by smooth trajectories with discontinuities that cluster over time are usually associated with exceptional events or structural breaks and call for the use of Markov-switching processes (Chen et al., 2019; Frühwirth-Schnatter, 2006).

The sudden and transient nature of these jumping patterns allow them to be characterized as exceptional risk amplification events affecting the intensity of social media users' concerns. Social amplification phenomena are known to occur on Web platforms, such as Twitter, especially for risk-related issues (Fellenor et al., 2018). We define social amplification as the sudden intensification of online communications containing a specific risk-related theme (e.g. a country). See Kasperson et al. (1988) for a theoretical framework of the social amplification of risks. The concerns of online communities can be amplified by salient events, online news coverage and other sociocultural factors affecting risk perception (Renn et al., 1992). Recently, Strekalova and Krieger (2017) found evidence of risk amplification in health-related risk debates occurring on Facebook. Similarly, Wang et al. (2016) analysed Twitter posts to identify the intensity of societal concerns regarding climate change-related risks.

Country risks are important in the explanation and prediction of market volatility indices (Ferreira & Gama, 2005). In particular, market operators are interested in identifying the specific contribution of each driver of public concerns in explaining the fluctuations of a volatility index to determine the country-risk spillover effects (Hoti, 2005). The measurement of country-risk perception may also empower government institutions with a new tool suitable for enacting their policy objectives, such as

issuing of new government bonds. Moreover, by disentangling the different components of country-related public concern and their relationship with volatility, it is possible to assess the relative contribution of rational and irrational behaviours in explaining volatility indices. Motivated by these facts, in this paper, we are concerned with the identification and disentanglement of country risks, which are then used to explain the fluctuations of a market volatility index, the VIX.

The contributions of this paper are manifold. First, we identify the timing, persistence and scale of social amplification events. Second, series-specific (country) effects are disentangled from common (global) contributions to intensity while controlling for exogenous factors that may contribute to explaining fluctuations in public concerns. Finally, the extracted intensities are used to highlight country-specific contributions to fluctuations in the VIX index.

The reminder of the paper is organized as follows. Section 2 presents the new statistical model for count time series data. The inferential procedure and algorithms are then described in Section 3. Next, the proposed methodology is used in Section 4 to study count data from Twitter posts that mention risk. Finally, Section 5 summarizes the main findings.

2 | A STATE SPACE MODEL FOR WEB COUNT DATA

Despite widespread interest in analysing risk perception and social amplification in online social media, statistical tools capable of exploiting specificity in the aforementioned data are limited. The peculiar features of count data obtained from social media, as discussed in the previous section, deserve special attention in the definition of a proper statistical model. We note that counts extracted from Web data and our dataset share similar characteristics, thus making the proposed statistical framework applicable to a wide range of empirical studies.

Modelling time series of counts pose several challenges, such as discreteness of the observations, temporal and cross-sectional dependence and over-dispersion. Despite renewed interest over the past decade (e.g. see Weiß, 2018), these issues still need to be resolved. Recent contributions include the dynamic Skellam model (Koopman et al., 2017) used for studying financial tick-by-tick data, Yang et al. (2015) study of health data with excess zeros, and the self-excited threshold Poisson model developed by Wang et al. (2014). Within the Bayesian approach, Park et al. (2011) used a zero-inflated Poisson model to study counts of user generated content in an on-line community.

In this Section, we propose a novel state space model for multivariate count time series which allows for: serial and cross-sectional correlation, sudden signal amplification events which cluster over time, and smooth dynamics. Computationally, the proposed model scales linearly in the cross-sectional dimension. We classify textual data from Twitter according to geographical markers (i.e. mention of the country in the text) to obtain a multivariate count time series $\{y_{1,t}, \ldots, y_{J,t}, z_t\}_t$ consisting of country-specific series $y_{j,t}$ for each $j = 1, \ldots, J$, as well as a global series z_t that stems from texts without a geographical reference to any of the selected countries. See Section 4 and the Supplement for further details on the construction of the dataset.

Starting from the count data illustrated in Figure 1, we aim to extract and disentangle components of the underlying dynamic intensity driving the observed counts: the public's concerns for country risk. As discussed in Section 1, the existence of unobserved public concerns driving Twitter posting activity suggests the use of a state space framework for count time series (Davis et al., 2016). This choice allows us to model data on their natural scale while preserving a direct interpretation of the components of the model. We follow this approach and assume that the country-specific observables, $y_{j,t}$, are Poisson distributed with a persistent, smooth latent intensity process, $x_{j,t}$, which represents country-risk concerns.

This also allows for positive autocorrelation and instantaneous cross-correlation, as found in the data (see Section 1 and the preliminary analysis in Section S.6.2. of the Supplement).

We account for persistent, smooth dynamics of positive-valued intensity $x_{j,t}$ by assuming a non-central Gamma distribution as the transition density of the process:

$$x_{j,t+1} \,|\, x_{j,t} \sim \text{NcGa}(\alpha_j, \beta_j x_{j,t}, \delta_j). \tag{1}$$

The non-central Gamma distribution is obtained as a Poisson mixture of Gamma. Hence if $x|z \sim \mathcal{G}a(a + z, c)$ with $z \sim \mathcal{P}oi(b)$, then $x \sim \text{NcGa}(a, b, c)$, where $\mathcal{G}a(a, c)$ denotes a Gamma distribution with shape a and scale c. It has density

$$P(x) = \exp\left(-\frac{x}{c}\right) \sum_{k=0}^{\infty} \frac{x^{a+k-1}}{c^{a+k}\Gamma(a+k)} \frac{b^k \exp(-b)}{k!}, \quad x \in \mathbb{R}_+, a > 0, b > 0, c > 0.$$

A non-central Gamma transition density defines an autoregressive Gamma process of order 1 (Gouriéroux & Jasiak, 2006), or ARG(1). The parameter α_j in Equation (3) governs the magnitude of innovation in the latent state process, whereas β_j, δ_j account for persistence. Gouriéroux and Jasiak (2006) proved the stationarity of the ARG(1) process if $\beta_j \delta_j < 1$ and derived the conditional mean and variance as $\mathbb{E}[x_{j,t+1} \,|\, x_{j,t}] = \delta_j \alpha_j + \beta_j \delta_j x_{j,t}$ and $\mathbb{V}[x_{j,t+1} \,|\, x_{j,t}] = \delta_j^2 \alpha_j + 2\beta_j \delta_j^2 x_{j,t}$. The choice of the ARG process for the latent states is motivated by its flexibility and the fact that it facilitates the structural interpretation of the latent states (public concern in our empirical application).

As discussed in Section 1, the sudden jumps magnifying the volume of observed counts in Figure 1 are due to social amplification phenomena. To avoid bias in the estimation of public concern, $x_{j,t}$, we introduce a positive multiplicative factor, $\xi_{j,t}$, thus obtaining the overall public concerns with amplification as

$$\tilde{x}_{j,t} = x_{j,t}(1 + \xi_{j,t}) = x_{j,t} + x_{j,t}\xi_{j,t}. \tag{2}$$

The multiplicative factor allows us to disentangle the contribution of transient social amplification events, $x_{j,t}\xi_{j,t}$, from the persistent evolution of country-risk concerns, $x_{j,t}$. Furthermore, because jumps in Twitter counts are persistent and cluster over time (see Section 1), we assume that country-specific social amplification is driven by an L-state hidden Markov chain $\{s_{j,t}\}_{t=1}^{T}$, such that $\xi_{j,t} = \xi_{j,s_{j,t}}$, where $\xi_{j,l}$ is a state-specific parameter. The transition probabilities of each chain j are assumed to be time-invariant and denoted by $\lambda_{j,l,k} = P(s_{j,t} = k \,|\, s_{j,t-1} = l)$, with transition matrix Λ_j. Finally, to account for the effects of exogenous variables on latent intensity, we include country-specific covariates, $\mathbf{v}_{j,t}$, in the analysis.

The resulting univariate model with local Markov-switching jumps has a state space representation:

$$\begin{aligned} y_{j,t} | x_{j,t}, \boldsymbol{\xi}_j, s_{j,t} &\sim \mathcal{P}oi(x_{j,t}(1 + \xi_{j,s_{j,t}}) + \exp(\mathbf{v}'_{j,t}\boldsymbol{\phi}_j)) \\ x_{j,t+1} | x_{j,t} &\sim \text{NcGa}(\alpha_j, \beta_j x_{j,t}, \delta_j) \end{aligned} \tag{3}$$

TABLE 1 Spearman's rank correlation among observed counts $(y_{j,t}, z_t)$ and latent local intensities, without amplification $(x_{j,t})$ and with amplification $(\tilde{x}_{j,t})$

Observations			**Latent multivariate**		**Latent univariate**	
$y_{US,t}, y_{UK,t}$	$y_{US,t}, z_t$	$y_{UK,t}, z_t$	$x_{US,t}, x_{UK,t}$	$\tilde{x}_{US,t}, \tilde{x}_{UK,t}$	$x_{US,t}, x_{UK,t}$	$\tilde{x}_{US,t}, \tilde{x}_{UK,t}$
0.257	0.447	0.387	0.066	0.071	0.234	0.249

To assess the univariate model's (3) ability to capture country-specific latent dynamics, we investigate the correlation structure of the observables counts and filtered latent intensities. Table 1 reports that the Spearman's rank correlation between the observed count time series is 0.257, whereas it is 0.234 between the latent intensities. This result indicates that the univariate model is unable to correctly extract the latent intensity at country level since the correlation between the observables is entirely transferred to the latent intensities. This is undesirable since we want latents to only capture the share of intensity specifically attributable to country-level features, removing the effect of common factors. This calls for the use of a multivariate framework that accounts for positive correlation among the observables while allowing the impact of co-movements and confounding factors to be disentangled from country-specific intensity.

To account for the cross-sectional dependence structure among the $y_{j,t}$, we distinguish country-specific public concerns from global ones, and include a common latent factor, w_t, in country-level intensity. The latent process w_t encodes Twitter users' perceptions of global (rather than country-specific) risks; thus, it is assumed to drive the temporal evolution of the global series, z_t. Note that including series z_t helps in isolating the impact of common risk factors and seasonal effects from the country-specific concerns. This specification leads to a multivariate model with local Markov-switching social amplification, whose conditional independence relationships are summarized by the DAG in Figure 2. The state space representation is

$$\begin{aligned} z_t|w_t &\sim \mathcal{P}oi(w_t + \exp(\mathbf{v}'_{z,t}\boldsymbol{\phi}_z)) \\ y_{j,t}|w_t, x_{j,t}, \boldsymbol{\xi}_j, s_{j,t} &\sim \mathcal{P}oi(w_t + x_{j,t}(1+\xi_{j,s_{j,t}}) + \exp(\mathbf{v}'_{j,t}\boldsymbol{\phi}_j)) \\ x_{j,t+1}|x_{j,t} &\sim \text{NcGa}(\alpha_j, \beta_j x_{j,t}, \delta_j) \\ w_{t+1}|w_t &\sim \text{NcGa}(\alpha_w, \beta_w w_t, \delta_w) \end{aligned} \tag{4}$$

We refer to the components of country-level intensity as follows: 'local' $x_{j,t}$, 'amplification' $x_{j,t}\xi_{j,s_{j,t}}$, 'global' w_t and 'covariates', $\exp(\mathbf{v}'_{j,t}\boldsymbol{\phi}_j)$.

The contribution of the proposed model for multivariate count time series is manifold. First, it accounts for series-specific and global latent intensities driving the observed counts, allowing to disentangle the impact of common factors from idiosyncratic latent dynamics. Jumps have been mainly studied in the context of real-valued time series (see Andersen et al., 2007, and references therein) using an additive specification. Instead, we allow for jumps in count time series through a series-specific multiplicative term which amplifies the effect of series-specific intensities. Despite their

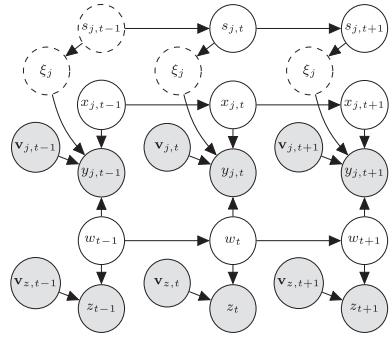

FIGURE 2 DAG of the model: observables (shaded grey circles), autoregressive components of the latent intensities and states (solid white circles) and amplification factors of latent intensity (dashed white nodes). Arrows' directions indicate the causal dependence relationships of the model

limited use in the literature (e.g. see Caporin et al., 2017), series-specific multiplicative jumps allow for a more flexible modelling as compared to the widespread additive specification. Finally, we do not make the temporal independence assumption for the jump terms and model them using a hidden Markov chain that allows to infer different states of excitation within the time series of counts.

Besides, the functional form of model (4) permits a direct structural interpretation of the latent variables and parameters as series-specific or global intensities, and amplification. This is granted by the assumption of an ARG process for the latent intensities. Differently, the commonly used lognormal specification Wang and Wang (2018); Heinen and Rengifo (2007), despite allowing the inclusion of similar terms in the intensity, would make the interpretation more difficult due to the nonlinear exponential link function.

Jørgensen et al. (1999) presented a state space for multivariate Poisson data where the observed counts are driven by a common Markov process. However, our framework is more general in two respects. First, we assume a different state transition that allows for both stationary and nonstationary processes. Second, our inferential procedure also holds for different observation densities. Recently, Chen et al. (2019) proposed a Markov-switching Poisson integer-valued GARCH model to account for consecutive zeros and time-varying volatility in count data. Since their model is univariate and observation driven, which means that the latent intensities are perfectly predictable, it is unable to account for features of the data highlighted in Section 1. Alternative approaches for multivariate count time series have been developed by Wang and Wang (2018), who modelled dependence by assuming a factor structure for latent intensities, and Heinen and Rengifo (2007), who exploited copulas to impose dependence among the observables while assuming a VARMA process for the dynamics of log intensities.

From a forecasting perspective, Berry and West (2020) used the decouple/recouple strategy to define a dynamic model for multivariate counts that is computationally scalable in the number of series. Conversely, our interest lies in identifying and disentangling the structural components of latent public concerns. This motivates the parametric assumptions underlying our model (4), which prevent the direct applicability of their decouple/recouple strategy.

Given the structural form of model (4), the likelihood function is a high-dimensional integral with no closed-form solution. Hence, we apply a data augmentation approach and introduce the latent state variables $s_{j,t}, x_{j,t}, w_t$ in the set of observations, thus obtaining the complete-data likelihood function

$$
\begin{aligned}
L(\mathbf{Y},\mathbf{Z},\mathbf{V},\mathbf{S},\mathbf{W},\mathbf{X}|\boldsymbol{\theta}) = & \prod_{t=1}^{T} \mathrm{NcGa}(\alpha_w, \beta_w w_{t-1}, \delta_w) \prod_{t=1}^{T}\prod_{j=1}^{J} \mathrm{NcGa}(\alpha_j, \beta_j x_{j,t-1}, \delta_j) \cdot \\
& \prod_{l=1}^{L}\prod_{j=1}^{J}\prod_{t\in\mathcal{T}_{j,l}} \frac{(w_t + x_{j,t}(1+\xi_{j,l}) + \exp(\mathbf{v}'_{j,t}\boldsymbol{\phi}_j))^{y_{j,t}}}{y_{j,t}!} \exp(-(w_t + x_{j,t}(1+\xi_{j,l}) + \exp(\mathbf{v}'_{j,t}\boldsymbol{\phi}_j))) \cdot \\
& \prod_{t=1}^{T} \frac{(w_t + \exp(\mathbf{v}'_{z,t}\boldsymbol{\phi}_z))^{z_t}}{z_t!} \exp(-(w_t + \exp(\mathbf{v}'_{z,t}\boldsymbol{\phi}_z))) \prod_{j=1}^{J}\prod_{l=1}^{L}\prod_{k=1}^{L} \lambda_{j,l,k}^{N_{lk}(\mathbf{S}_j)} p(\mathbf{s}_{j,0}|\Lambda_j),
\end{aligned}
\tag{5}
$$

where $\mathcal{T}_{j,l} = \{t: s_{j,t} = l\}$ and $N_{lk}(\mathbf{S}_j) = \#\{s_{j,t-1} = l, s_{j,t} = k\}$ are the number of transitions from state l to state k of the j-th chain $\mathbf{S}_j$. Motivated by our interest in identifying of social amplification phenomena in data, we assume $L = 2$ and impose $\xi_{j,1} = 0$ for every j. This leads to identifying regime 1 as having no jumps and regime 2 as the social amplification regime. However, the model specification is general and allows for any number of regimes $L > 1$ that can be identified using the constraint $\xi_{j,1} < \ldots < \xi_{j,L}$.

3 | BAYESIAN INFERENCE

3.1 | Prior specification

In this paper, we adopt a Bayesian approach that provides a simple way to introduce regularization via the specification of appropriate prior distributions and permits a flexible modelling of jumps through hierarchical priors. The parameters α_j and δ_j (similarly, α_w and δ_w) of the non-central Gamma distribution in (4) are not separately identifiable, thus requiring the introduction of constraints for their estimation. To solve the identification issue, we introduce a soft constraint by specifying a truncated prior distribution for δ_j, δ_w, which bounds them away from zero, and then test the robustness of the results for various truncation levels. In all the analyses performed, we find that the parameters δ_j, δ_w are never stuck at the lower bound, thus providing evidence that the proposed constraint is non-binding. See the Supplement for further details regarding the proposed truncation scheme and values of the hyper-parameters.

We assume a flexible hierarchical prior for the country-specific social amplification factor, $\xi_{j,2}$. The overall prior structure for each $j = 1, \ldots, J$ is

$$\begin{aligned}
&\eta_j \sim \mathcal{G}a(a_\eta, b_\eta) \qquad \gamma_j|\eta_j \propto \frac{a_\gamma^{\gamma_j - 1}\eta_j^{\gamma_j c_\gamma}}{\Gamma(\gamma_j)^{b_\gamma}} \qquad \xi_{j,2}|\gamma_j, \eta_j \sim \mathcal{G}a(\gamma_j, 1/\eta_j),\\
&\boldsymbol{\phi}_j \sim \mathcal{N}(\underline{\boldsymbol{\phi}}, \underline{\Sigma}) \qquad \boldsymbol{\phi}_z \sim \mathcal{N}(\underline{\boldsymbol{\phi}}_z, \underline{\Sigma}_z)\lambda_{j,l} \sim \mathcal{D}ir(\underline{\lambda}_j)\\
&\alpha_j \sim \mathcal{G}a(a_\alpha, b_\alpha) \qquad \beta_j \sim \mathcal{G}a(a_\beta, b_\beta) \qquad \delta_j \sim T\mathcal{G}a(a_\delta, b_\delta; S_\tau),\\
&\alpha_w \sim \mathcal{G}a(a_{\alpha_w}, b_{\alpha_w}) \qquad \beta_w \sim \mathcal{G}a(a_{\beta_w}, b_{\beta_w}) \qquad \delta_w \sim T\mathcal{G}a(a_{\delta_w}, b_{\delta_w}; S_{\tau_w}).
\end{aligned} \tag{6}$$

The notation $T\mathcal{G}a(a, b; \tau)$ stands for a Gamma distribution truncated on the interval $S_\tau = (0, \tau) \subset \mathbb{R}_+$, parametrized by τ. Finally, the unnormalized prior distribution for γ_j is conjugated for the shape parameter of a Gamma distribution (see Llera & Beckmann, 2016) with positive real hyper-parameters $a_\gamma, b_\gamma, c_\gamma$. The conjugacy property allows for efficient posterior sampling via the inverse transform method without the need for a tuning parameter. Figure 3 reports the directed acyclic graph (DAG) of the model and prior structure.

3.2 | Posterior approximation

The joint posterior distribution of parameters and latent variables is

$$P(\boldsymbol{\theta}, \mathbf{S}, \mathbf{W}, \mathbf{X} \,|\, \mathbf{Y}, \mathbf{Z}, \mathbf{V}) = P(\boldsymbol{\theta}) L(\mathbf{Y}, \mathbf{Z}, \mathbf{V}, \mathbf{S}, \mathbf{W}, \mathbf{X} \,|\, \boldsymbol{\theta}).$$

Since this distribution is not tractable, we develop a MCMC algorithm to generate random draws from the posterior distribution and approximate all posterior quantities of interest. We refer to Appendix A and the Supplement for the detailed derivation of posterior distributions.

Estimating the trajectory of latent intensity is known as the filtering problem. For special classes of state space models, such as linear Gaussian state space models, the filtering distribution is known in closed form, and efficient algorithms are available (e.g. the Kalman and Hamilton filters). In more general frameworks such as Equation (4), when no closed-form expressions are available, the filtering problem may be solved by resorting to simulation-based methods, such as particle filters. We follow this strategy and rely on the selection/mutation (SM) algorithm (e.g. see Cappé et al., 2005) to obtain

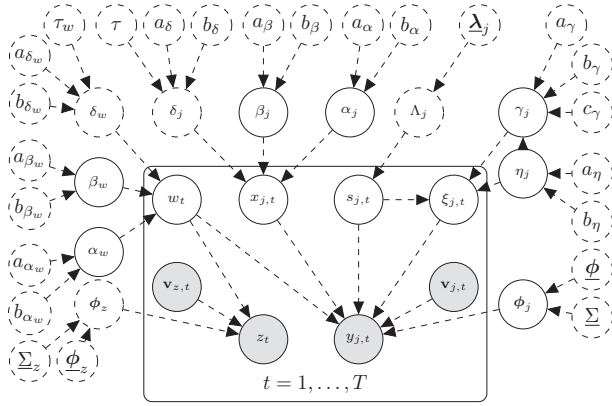

FIGURE 3 DAG of model (4). It exhibits the hierarchical structure of priors and related hyper-parameters: observables (shaded grey circles), latent parameters (solid white circles) and fixed hyper-parameters (dashed white circles). Arrows' directions indicate the causal dependence relationships of the model

filtered estimates of the paths of latent states w_t and $x_{j,t}$. The SM algorithm addresses the degeneracy issue by selecting, at each point in time, a set of particles based on the associated weights and then simulating an independent extension for each selected trajectory. The static parameters governing the dynamics and amplification factors are sampled using an adaptive Metropolis–Hastings (aRWMH) step (Atchadé & Rosenthal, 2005). Finally, hyper-parameters are directly sampled from their corresponding posterior full conditional distributions.

We use the convention $\mathbf{X}_j = \{x_{j,t}: t = 1, \ldots, T\}$ and $\mathbf{X} = \{\mathbf{X}_j: j = 1, \ldots, J\}$ and similarly for the other variables and parameters. The MCMC algorithm is articulated in two main blocks, drawing the path of global intensity (w_t) and country-specific components ($x_{j,t}$, $s_{j,t}$, $\xi_{j,s_{j,t}}$), as follows:

1. Sample the global latent intensity and the static parameters:
   - (1a) Sample the latents $\mathbf{W}$ conditionally on $(\mathbf{Y}, \mathbf{Z}, \mathbf{V}, \alpha_w, \beta_w, \delta_w, \mathbf{X}, \boldsymbol{\xi}, \mathbf{S})$ using a particle filter with the SM algorithm;
   - (1b) Sample the static parameters of the ARG process for w_t:
     - sample α_w from $P(\alpha_w \mid \mathbf{W}, \beta_w, \delta_w)$ via aRWMH;
     - sample β_w from $P(\beta_w \mid \mathbf{W}, \alpha_w, \delta_w)$ via aRWMH;
     - sample δ_w from $P(\delta_w \mid \mathbf{W}, \alpha_w, \beta_w)$ via aRWMH;
   - (1c) Sample the coefficients $\boldsymbol{\phi}_z$ from $P(\boldsymbol{\phi}_z \mid \mathbf{Z}, \mathbf{V}_z, \mathbf{W})$ via aRWMH;
2. Independently for each $j = 1, \ldots, J$, sample the country-specific latent intensity, hidden Markov chain and static parameters:
   - (2a) Sample the latents $\mathbf{X}_j$ conditionally on $(\mathbf{Y}_j, \mathbf{V}_j, \alpha_j, \beta_j, \delta_j, \boldsymbol{\xi}_j, \mathbf{W}, \mathbf{S}_j)$ using a particle filter with the SM algorithm;
   - (2b) Sample the hidden chain $\mathbf{S}_j$ conditionally on $(\mathbf{Y}_j, \mathbf{V}_j, \mathbf{W}, \mathbf{X}_j, \boldsymbol{\xi}_j, \Lambda_j)$ using the FFBS algorithm;
   - (2c) Sample the rows of transition matrix Λ_j from $P(\lambda_{j,l} \mid \mathbf{S}_j)$ for $l = 1,2$;
   - (2d) Sample, in block, the parameters associated with the jump terms:
     - sample η_j from $P(\eta_j \mid \boldsymbol{\xi}_j, \gamma_j)$;
     - sample γ_j from $P(\gamma_j \mid \boldsymbol{\xi}_j, \eta_j)$ via the inverse transform method;
     - sample the amplification $\xi_{j,2}$ from $P(\xi_{j,2} \mid \mathbf{Y}_j, \mathbf{X}_j, \mathbf{W}, \gamma_j, \eta_j, \mathbf{S}_j)$ via aRWMH;
   - (2e) Sample the static parameters of the ARG process for $x_{j,t}$:
     - sample α_j from $P(\alpha_j \mid \mathbf{X}_j, \beta_j, \delta_j)$ via aRWMH;
     - sample β_j from $P(\beta_j \mid \mathbf{X}_j, \alpha_j, \delta_j)$ via aRWMH;

- sample δ_j from $P(\delta_j \mid \mathbf{X}_j, \alpha_j, \beta_j)$ via aRWMH;

(2f) Sample the coefficients $\boldsymbol{\phi}_j$ from $P(\boldsymbol{\phi}_j \mid \mathbf{Y}_j, \mathbf{V}_j, \mathbf{X}_j, \mathbf{S}_j, \boldsymbol{\xi}_j)$ via aRWMH.

We tested the sampler's performance in simulated experiments and compared it to a lognormal specification for latent intensities $w_t, x_{j,t}$. The results suggest that our framework is better suited for minimizing the risk of miss-classifying ordinary fluctuations of the observables as social amplification events, which is one of the objectives in our application. See the Supplement for further details. Note that model (4) and the proposed sampler can be easily modified to analyse binary or integer-valued time series since the observational densities in Equation (4) only affect the distribution to be approximated by the particle filter.

4 | EMPIRICAL APPLICATION

In this application we model the public concern among Twitter users regarding country-related risks in the UK and US. According to Chung (2011), '*easy access to and efficient sharing of risk information [...] could accelerate the intensity as well as the speed of social attention to risk issues*'. Accordingly, we apply model (4) to count data generated from Twitter posts mentioning risk to extract the intensity of country-risk concerns (e.g. political, macroeconomic, financial or environmental) and identify risk amplification events.

The period of investigation, which ranges from 1 April 2016 to 1 March 2017, is characterized by the UK's European Union membership referendum (23 June 2016), the US presidential elections (8 November 2016), and their aftermath, including the Brexit negotiations and the Trump administration taking office. Referendums and presidential elections are major political events, so they receive broad media coverage and are characterized by intense speculation on their outcome and consequences. Since these events can potentially generate public concern regarding their implications at the national and international level, this dataset provides an ideal setting for disentangling the different components of country risks and evaluating their impact on the volatility of financial markets. Figure 4 summarizes the workflow of the empirical application.

4.1 | Data collection

Tweets containing the word *risk* were collected by programmatically querying the Search API of Twitter (Makice, 2009). Further details about the data strategy and query parameters are included in Section S.6.1 of the Supplement. For each day of the period under investigation, we count the number of Twitter posts (considering both tweets and retweets) matching specific country dictionary

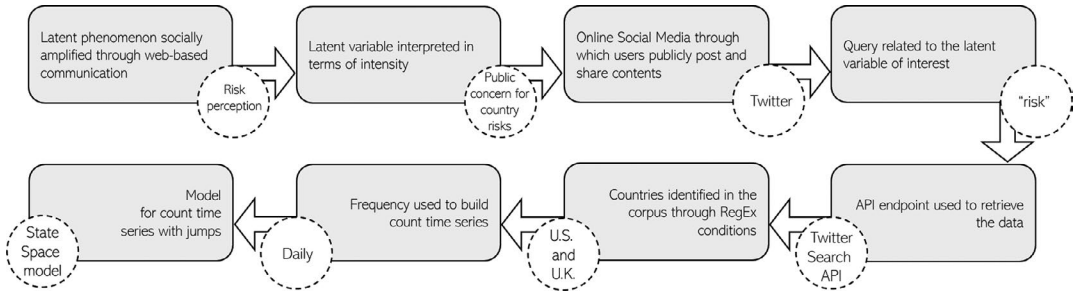

FIGURE 4 Flow chart of the implemented research procedure

conditions (see RegEx Tables 17 and 18 in the Supplement), which are used to identify tweets referring to country-specific risks affecting either the US or UK. By doing so, we obtain the count time series $y_{US,t}$ and $y_{UK,t}$. Note that filtering risk data by country dictionaries without focusing on specific subtopics offers a broad lens to view all the (potentially time-varying) salient dimensions of country-risk perception. Conversely, some of these dimensions may be missed when using topical dictionaries since they cannot provide, *a-priori*, an exhaustive list of all the drivers of country risks.

Moreover, while country dictionary conditions (e.g. country names and acronyms) are exhaustive, unambiguous and stable over time, topic dictionary conditions with the same qualities are much more difficult to construct. Similarly, we create the global series z_t by counting tweets about risk(s) that do not refer directly to the US or UK (i.e. tweets that do not match either country's dictionary conditions). Tweets counted the global series may refer to risks related to policies (e.g. Trump's foreign policy), events (e.g. the Brexit) and international factors (e.g. protectionism and international trade wars) which may affect both countries.

To control for the possible effect of breaking news and scheduled events that are expected to affect the volatility of financial markets, we also include (i) the share of newspaper articles containing the word *risk*, at both the country and global level, and (ii) the country-specific number of expected high-volatility events as covariates. See Section S.6.1 of the Supplement for further details.

4.2 | Results

Our data consist of $T = 334$ daily observations for $J = 2$ countries (the US and UK), and a global series. The count series for the UK and US are rescaled by a factor of 0.1 and the global series by a factor of 0.001. We run the Gibbs sampler for 20,000 iterations after discarding the first 4000 as burn-in.

Figure 5 shows the original data and the estimated latent intensity, distinguishing the persistent country-specific part (red line) from the estimated social amplification phenomena (red stars). The sampler identifies jumps that correspond to days characterized by the highest variation rates in the observations. In the first 4 months, which include the Brexit referendum, the path of the persistent component for the UK experienced a tumultuous period with higher average values compared to the following months. For the US, latent intensity appears rather stable, with some turmoil from mid-July to the US elections in November, as Donald Trump's polling against Hillary Clinton rose steadility (Bovet et al., 2018), as well as when he took office in late January 2017.

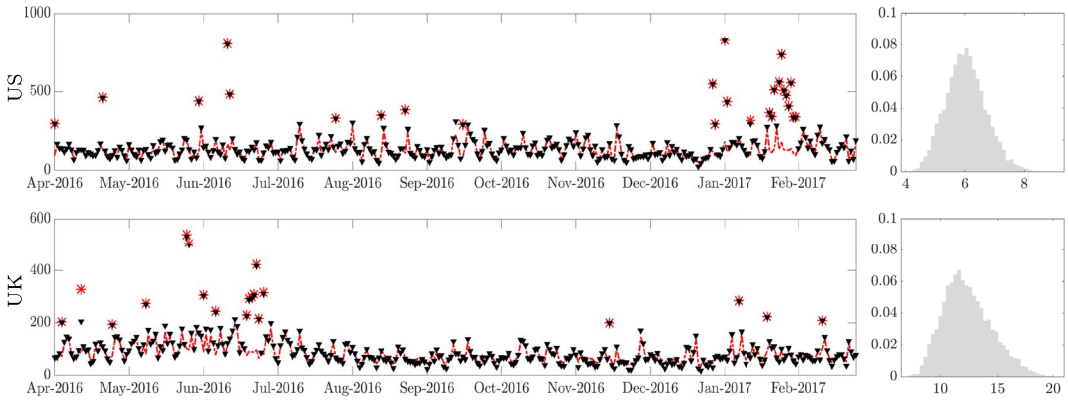

FIGURE 5 *Left:* observed data (black triangles) and the posterior mean of filtered states (dashed red lines) and jumps (red stars). *Right:* posterior distribution of the social amplification factor [Colour figure can be viewed at wileyonlinelibrary.com]

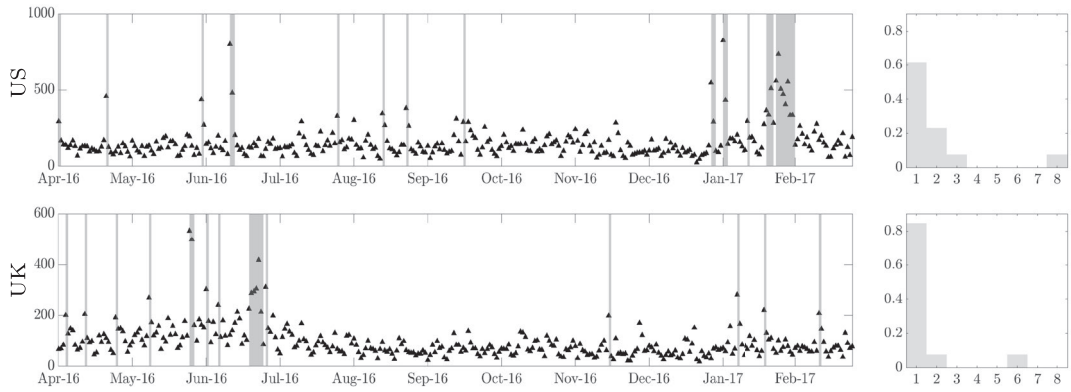

FIGURE 6 *Left:* observation (black triangles) and estimated hidden chain $s_{j,t}$ or the US and UK: regime 2 (grey bars) and regime 1 (white bars). *Right:* distribution of the duration of the social amplification regime (regime 2)

The days in which social amplification phenomena occur are reported as grey shades in Figure 6, along with the distribution of their duration. The different country-level results corroborate the interpretation of social amplification as a local and unsynchronized phenomenon. In particular, most UK events occur before and right after the Brexit referendum, whereas social amplification in the US is more concentrated when Donald Trump took office. In particular, we find that during the first 3 months of the sample, corresponding to the period immediately prior to the EU referendum, which was characterized by the Brexit victory and resignation of PM David Cameron, social amplification of risk with respect to the UK was frequent.

However, we find that events of social amplification are relatively more frequent for the US (i) at the beginning of June 2016, (ii) in August 2016 and (iii) from late December 2016 to February 2017. As aforementioned, interval (ii) was characterized by temporary reductions in the gap between Clinton and Trump in the polls, whereas the interval (iii) follows Donald Trump taking office as the President of the United States, and corresponds to the commencement of Trump's 'America First' domestic and foreign policy.

Moreover, from the right-hand column of Figures 5 and 6, we find that the average duration of amplification events in the UK is lower compared to the US, but their amplitude is higher in relative terms, capturing the fact that social amplification in the UK is more exceptional and intense compared to the US. This provides evidence in favour of the interpretation of social amplification as a transient phenomenon, which rapidly dissipates without having any persistent effect on the public's perception of risk.

The composition of total intensity for the countries under investigation is shown in Figure 7, which reports the daily share of each component over the total. Overall, we find that in both countries, (i) the share of local intensity, $x_{j,t}$, evolves quite smoothly over time, and (ii) social amplification, when occurring, accounts for about 70% of the total intensity in both countries. These results provide evidence of time variation in the drivers of country-risk perception, highlighting the role of social amplification in periods of extraordinary political events (e.g. referendums, elections, new presidents taking office), and the relative importance of the local and global components of risk perception on ordinary days. In particular, since August 2016 the weight of country-risk concerns, $x_{j,t}$ (dark grey), is significantly higher for the US than the UK, whereas the latter series is found to be mainly driven by the global term, w_t (light grey). This may be a consequence of UK-related concerns experiencing a drop a few weeks after the Brexit referendum, aligning with the level of non-country-related risks. This drop may also be due to a switch in the nature of Brexit-related risks after the referendum, which started being perceived as global risks. Overall, these results suggest that for the UK, local factors are the most influential driver of public concern only in the first part of the sample, whereas the forces driving the intensity of US concerns are more evenly distributed, with the local factor playing the major role.

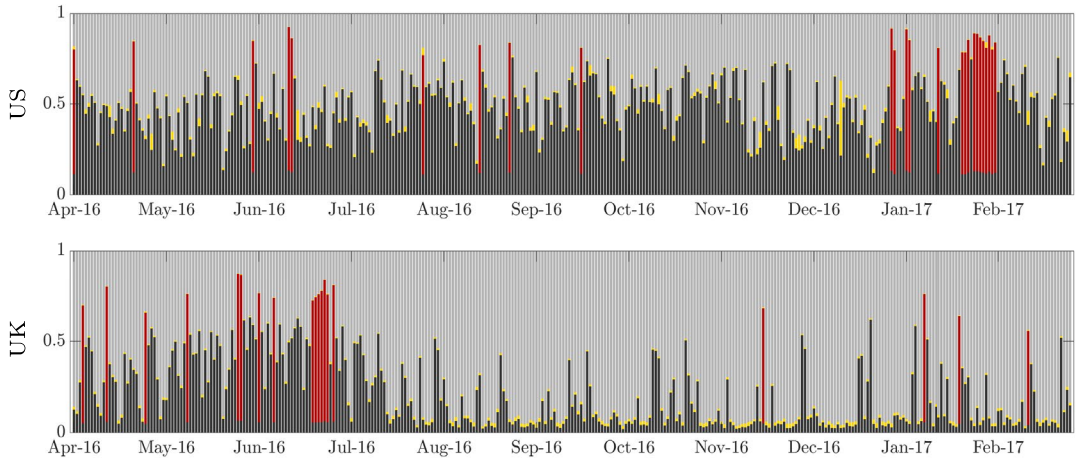

FIGURE 7 Daily composition of intensities in percentage: local $x_{j,t}$ (dark grey shade), amplification $x_{j,t}\xi_{j,2}$ (red), covariates $\exp(\boldsymbol{\phi}_j'\mathbf{v}_{j,t})$ (gold), and global w_t (light grey) [Colour figure can be viewed at wileyonlinelibrary.com]

To deepen the analysis of social amplification phenomena, Table 2 reports the three most frequent words written during these periods (i.e. days t such that $s_{j,t} = 2$). Interestingly, we find that the these words change over time, providing evidence that social amplification is not related to a specific event or topic; instead, it rather captures swings in posting activities regarding all topical dimensions of country risks.

For example, the jump in the US series in May 2016 refers to concerns related to the possible spread of the Zika virus in the US. Interestingly, the UK-risk perception jumped in January 2017 in response to Donald Trump taking office, thus showing that local events may have significant ripples in the public's concerns regarding other countries. In summary, the results presented in Table 2 support the hypothesis that on days with social amplification, people are more intensively communicating their concerns regarding major political events. Hence the public concerns regarding these country risks may be amplified through social media.

We extend our analysis to include Germany and China and obtained results similar to those for the US and UK. Furthermore, we used topic-based classification of tweets to investigate political and economical risk dimensions. Interestingly, we find that the social amplification of the intensity of economic risks occurred close to the Brexit referendum, whereas Donald Trump taking office is associated with a political risk social amplification phenomenon. We refer the reader to Section S.6 of the Supplement for further details.

Motivated by the long-standing debate on the relationship between risk perception and financial markets, we now investigate the performance of the extracted intensity of country-risk concerns in explaining fluctuations of a volatility index, the VIX. This analysis offers new ways to evaluate the impact of social media amplification phenomena on market volatility. Figure 8 plots the total intensity of country-risk concerns for the US and UK (blue and red) against the VIX index (black).

To explore this relationship further, we exploit the decomposition of country risks into the persistent and social amplification components (see Equation (2)) to assess the contribution of each factor in explaining fluctuations in the VIX. Table 3 reports the estimation results of the linear regression:

$$\Delta VIX_t = \alpha + \boldsymbol{\beta}_1'\Delta\mathbf{f}_{1,t} + \ldots + \boldsymbol{\beta}_J'\Delta\mathbf{f}_{J,t} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(0,\sigma^2), \tag{7}$$

where we assume a non-informative prior and use four alternative specifications for the vector of covariates, $\mathbf{f}_{j,t}$: (i) local intensity with amplification, $\tilde{x}_{j,t}$, (ii) jumps, $x_{j,t}\xi_{j,s_{j,t}}$, (iii) local intensity, $x_{j,t}$ and (iv) local intensity and amplification, $(x_{j,t}, x_{j,t}\xi_{j,s_{j,t}})'$. We consider three models: two including only a single country

TABLE 2 Dates of occurrence of social amplification (regime 2) and the three most frequently used words in posts of the corresponding days (excluding retweets and tweet duplicates)

US			UK		
From	**To**	**Top 3 words**	**From**	**To**	**Top 3 words**
1 April 2016	1 April 2016	jobs, ban, lives	4 April 2016	4 April 2016	cfos, brexit, country
21 April 2016	21 April 2016	judges, liberal, forecast	12 April 2016	12 April 2016	forecast, tata, steel
31 May 2016	31 May 2016	zika, researchers, infection	25 April 2016	25 April 2016	terror, new, high
12 June 2016	13 June 2016	synthetic, drugs, pose	9 May 2016	9 May 2016	forecast, may, pollution
27 July 2016	27 July 2016	data, security, ups	26 May 2016	27 May 2016	students, referendum, vote
15 August 2016	15 August 2016	million, increase, children	2 June 2016	2 June 2016	vote, global, growth
25 August 2016	25 August 2016	just, blood, using	7 June 2016	7 June 2016	forecast, big, brexit
18 September 2016	18 September 2016	ups, early, heart	20 June 2016	25 June 2016	brexit, forecast, exit
31 December 2016	1 January 2017	obama, ideals, trump	27 June 2016	27 June 2016	data, security, fire
5 January 2017	6 January 2017	trump, data, lives	18 November 2016	18 November 2016	bankers, warn, exodus
16 January 2017	16 January 2017	trump, jobs, man	11 January 2017	11 January 2017	trump, details, users
24 January 2017	26 January 2017	fbi, hillary, grave	23 January 2017	23 January 2017	free, people, high
28 January 2017	4 February 2017	security, grid, utility	15 February 2017	15 February 2017	legal, sector, people

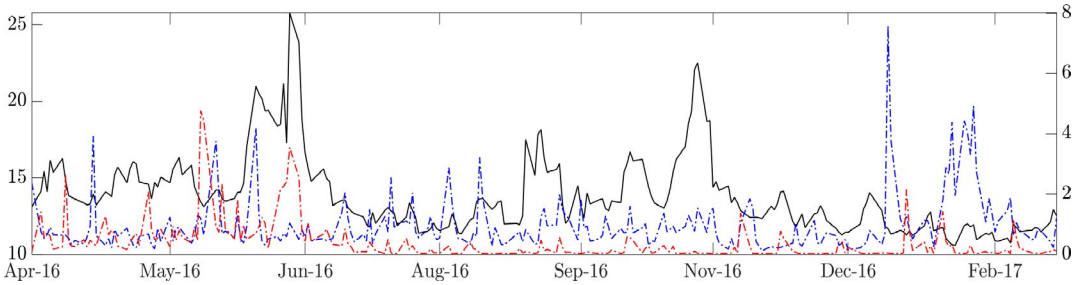

FIGURE 8 VIX (solid black line, left axis) and local intensities with amplification (right axis): the US (dashed blue line) and the UK (dashed red line). For visualization purposes, local intensities have been rescaled by a factor of 0.01 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Bayesian linear regression of VIX on *local intensity with amplification*, $\Delta\tilde{x}_{j,t}$ (top-left); *local intensity*, $\Delta x_{j,t}$ (top-right); *jumps*, $\Delta x_{j,t}\xi_{j,s_{j,t}}$ (bottom-left); *local intensity*, $\Delta x_{j,t}$, and *jump*, $\Delta x_{j,t}\xi_{j,s_{j,t}}$ (bottom-right). Standard deviations of the posterior distributions are in parentheses. Coefficients with posterior credible intervals not containing zero are shaded in grey

	(a)	(b)	(c)		(a)	(b)	(c)
	US only	**UK only**	**US, UK**		**US only**	**UK only**	**US, UK**
const	−0.008	−0.007	−0.005	*const*	−0.008	−0.006	−0.008
	(0.081)	(0.081)	(0.080)		(0.081)	(0.081)	(0.081)
$\Delta\tilde{x}_{US}$	0.162		0.171	Δx_{US}	0.153		0.157
	(0.083)		(0.082)		(0.155)		(0.153)
$\Delta\tilde{x}_{UK}$		0.232	0.246	Δx_{UK}		0.337	0.324
		(0.130)	(0.128)			(0.318)	(0.320)
DIC	744.814	745.544	742.997	*DIC*	747.778	747.635	748.673
const	−0.007	−0.006	−0.005	*const*	−0.007	−0.005	−0.005
	(0.082)	(0.082)	(0.081)		(0.082)	(0.080)	(0.081)
$\Delta x_{US}\xi_{US}$	0.121		0.136	Δx_{US}	0.221		0.202
	(0.083)		(0.083)		(0.157)		(0.157)
$\Delta x_{UK}\xi_{UK}$		0.156	0.180	$\Delta x_{US}\xi_{US}$	0.151		0.159
		(0.121)	(0.122)		(0.086)		(0.087)
				Δx_{UK}		0.533	0.500
						(0.342)	(0.337)
				$\Delta x_{UK}\xi_{UK}$		0.226	0.237
						(0.130)	(0.129)
DIC	746.655	747.019	746.354	*DIC*	746.580	746.514	746.283

(a-b) and another with both countries (c). The best model for explaining fluctuations in the VIX includes the total risk intensity for both countries. We find that including risk intensity for both countries tends to improve the model's performance, thus suggesting the existence of cross-country risk spillover effects. Moreover, by disaggregating local intensity from the jump component, the results show that neither component in isolation has a better performance than their combination. If we also consider the results of the full model (bottom-right part of Table 3), this may be interpreted as evidence that both components are relevant, with social amplification playing a major role.

This analysis shows that the flexible decomposition of country-risk intensity enabled by our framework allows us to investigate the impact of its components (and combinations of them) on a market index, such as the VIX.

5 | CONCLUSIONS

In this paper, we investigated the relation between country-risk perception and financial markets using a measure of public concern at the country level extracted from online social media data. Count time series obtained from Web data pose several challenges, including the coexistence of jumps and smooth dynamics, with cross-sectional effects and common dynamics. To address these challenges,

we designed a novel state space framework for multivariate time series of counts that is able to account for both of these features. The total intensity of each country series is given by the combination of a series-specific autoregressive term driving the persistent smooth dynamics, an idiosyncratic jump term that accounts for social amplification, and a global term that captures co-movements.

The proposed method is applied to count time series extracted from Twitter posts about UK and US country risks. Using our model, we were able to identify periods of social amplification and relate them to specific events. Finally, to highlight the value of extracted components, we investigated their impact on the VIX and found evidence of country-risk spillover effects, driven primarily by social amplification phenomena. The proposed methodology is general and has a wide range of applications in other scientific fields where count time series are popular, such as biomedical and epidemiological applications.

ACKNOWLEDGEMENTS

We are grateful to the Joint Editor, Jouni Kuha, the Associate Editor and two anonymous reviewers for many insightful comments that helped improve and clarify this manuscript. Carlo Santagiustina acknowledges financial support from the European Union ODYCCEUS Horizon 2020 project, grant agreement number 732942. Matteo Iacopini acknowledges financial support from the EU Horizon 2020 programme under the Marie Sklodowska-Curie scheme (grant agreement no. 887220).

ORCID

Matteo Iacopini http://orcid.org/0000-0002-3551-4891
Carlo R.M.A. Santagiustina https://orcid.org/0000-0003-3253-1263

REFERENCES

Aktekin, T., Polson, N. & Soyer, R. (2018) Sequential Bayesian analysis of multivariate count data. *Bayesian Analysis*, 13(2), 385–409.

Andersen, T.G., Bollerslev, T. & Diebold, F.X. (2007) Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4), 701–720.

Atchadé, Y.F. & Rosenthal, J.S. (2005) On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5), 815–828.

Berry, L.R. & West, M. (2020) Bayesian forecasting of many count-valued time series. *Journal of Business & Economic Statistics*, 38(4), 872–887.

Bialkowski, J., Gottschalk, K. & Wisniewski, T.P. (2008) Stock market volatility around national elections. *Journal of Banking & Finance*, 32(9), 1941–1953.

Bovet, A., Morone, F. & Makse, H.A. (2018) Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton versus Donald Trump. *Scientific Reports*, 8(1), 8673.

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

Caporin, M., Rossi, E. & de Magistris, P.S. (2017) Chasing volatility: A persistent multiplicative error model with jumps. *Journal of Econometrics*, 198(1), 122–145.

Cappé, O., Moulines, E. & Rydén, T. (2005) *Inference in hidden Markov models*. Berlin: Springer.

Chen, C.W., Khamthong, K. & Lee, S. (2019) Markov switching integer-valued generalized auto-regressive conditional heteroscedastic models for dengue counts. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(4), 963–983.

Chung, I.J. (2011) Social amplification of risk in the internet environment. *Risk Analysis: An International Journal*, 31(12), 1883–1896.

D'Amuri, F. & Marcucci, J. (2017) The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816.

Davis, R.A., Holan, S.H., Lund, R. & Ravishanker, N. (2016) *Handbook of discrete-valued time series*. Boca Raton: CRC Press.

Durbin, J. & Koopman, S.J. (2012) *Time series analysis by state space methods*. Oxford: Oxford University Press.
Einav, L. & Levin, J. (2014) The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1–24.
Fellenor, J., Barnett, J., Potter, C., Urquhart, J., Mumford, J. & Quine, C. (2018) The social amplification of risk on Twitter: the case of ash dieback disease in the United Kingdom. *Journal of Risk Research*, 21(10), 1163–1183.
Ferreira, M.A. & Gama, P.M. (2005) Have world, country, and industry risks changed over time? An investigation of the volatility of developed stock markets. *Journal of Financial and Quantitative Analysis*, 40, 195–222.
Frühwirth-Schnatter, S. (2006) *Finite mixture and Markov switching models*. Berlin: Springer Science & Business Media.
Gouriéroux, C. & Jasiak, J. (2006) Autoregressive Gamma processes. *Journal of Forecasting*, 25(2), 129–152.
Hassan, M.K., Maroney, N.C., El-Sady, H.M. & Telfah, A. (2003) Country risk and stock market volatility, predictability, and diversification in the middle east and Africa. *Economic Systems*, 27(1), 63–82.
Heinen, A. & Rengifo, E. (2007) Multivariate autoregressive modeling of time series count data using copulas. *Journal of Empirical Finance*, 14(4), 564–583.
Hoti, S. (2005) Modelling country spillover effects in country risk ratings. *Emerging Markets Review*, 6(4), 324–345.
Howell, L.D. (2011) *International country risk guide methodology*. East Syracuse, NY: PRS Group.
Huber, J., Palan, S. & Zeisberger, S. (2019) Does investor risk perception drive asset prices in markets? Experimental evidence. *Journal of Banking & Finance*, 108, 105635.
Jørgensen, B., Lundbye-Christensen, S., Song, P.-K. & Sun, L. (1999) A state space model for multivariate longitudinal count data. *Biometrika*, 86(1), 169–181.
Kasperson, R.E., Renn, O., Slovic, P., Brown, H.S., Emel, J., Goble, R. et al. (1988) The social amplification of risk: a conceptual framework. *Risk Analysis*, 8(2), 177–187.
Koopman, S.J., Lit, R. & Lucas, A. (2017) Intraday stochastic volatility in discrete price changes: the dynamic Skellam model. *Journal of the American Statistical Association*, 112(520), 1490–1503.
Llera, A. & Beckmann, C. (2016) Bayesian estimators of the Gamma distribution. *arXiv preprint arXiv:1607.03302*.
Makice, K. (2009) *Twitter API: Up and running learn how to build applications with the Twitter API*. Sebastopol: O'Reilly Media, Inc.
Park, Y.-H., Park, C.H. & Ghosh, P. (2011) Modelling member behaviour in on-line user-generated content sites: a semiparametric Bayesian approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(4), 1051–1069.
Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. & Mozetič, I. (2015) The effects of Twitter sentiment on stock price returns. *PloS One*, 10(9), e0138441.
Renn, O., Burns, W.J., Kasperson, J.X. Kasperson, R.E. & Slovic, P. (1992) The social amplification of risk: theoretical foundations and empirical applications. *Journal of Social Issues*, 48(4), 137–160.
Rogers, R. (2013) Debanalizing Twitter: The transformation of an object of study. In: *Proceedings of the 5th Annual ACM Web Science Conference*, pp. 356–365. ACM.
Strekalova, Y.A. & Krieger, J.L. (2017) Beyond words: amplification of cancer risk communication on social media. *Journal of Health Communication*, 22(10), 849–857.
Varian, H.R. (2014) Big data: new tricks for econometrics. *Journal of Economic Perspectives*, 28(2), 3–28.
Wang, F. & Wang, H. (2018) Modelling non-stationary multivariate time series of counts via common factors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4), 769–791.
Wang, C., Liu, H., Yao, J.-F., Davis, R.A. & Li, W.K. (2014) Self-excited threshold Poisson autoregression. *Journal of the American Statistical Association*, 109(506), 777–787.
Wang, Z., Ye, X. & Tsou, M.-H. (2016) Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards*, 83(1), 523–540.
Weiß, C.H. (2018) *An introduction to discrete-valued time series*. Hoboken: John Wiley & Sons.
World Economic Forum. (2017) Global risks report 2017. Geneva: world economic forum. Available at http://www3.weforum.org/docs/GRR17_Report_web.pdf.
Yang, M., Cavanaugh, J.E. & Zamba, G.K. (2015) State-space models for count time series with excess zeros. *Statistical Modelling*, 15(1), 70–90.
Zaman, T., Fox, E.B. & Bradlow, E.T. (2014) A Bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3), 1583–1611.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Iacopini M, Santagiustina CR. Filtering the intensity of public concern from social media count data with jumps. *J R Stat Soc Series A*. 2021;184:1283–1302. https://doi.org/10.1111/rssa.12704

APPENDIX A

POSTERIOR DISTRIBUTIONS

This section reports the posterior distributions of the parameter in model (4). See the Supplement for the detailed computations. The adaptive random walk Metropolis–Hastings (aRWMH) steps have been designed following Atchadé and Rosenthal (2005), with an asymptotic acceptance rate of $\bar{r} = 0.30$.

A.1 Sampling latents: W, $\mathbf{X}_j$, and $\mathbf{S}_j$

The trajectory of the global latent intensity $\mathbf{W}$ is estimated conditionally on $(\mathbf{Y}, \mathbf{Z}, \mathbf{V}, \boldsymbol{\phi}_z, \alpha_w, \beta_w, \delta_w, \mathbf{X}, \boldsymbol{\xi}, \mathbf{S})$. Since the filtered distribution from the state space in Equation (4) is not available in closed form, we approximate it using a particle filter based on the Selection/Mutation (SM) algorithm (Cappé et al., 2005). For analogous reasons, the SISR method is used to generate the trajectory of the country-specific autoregressive latent intensity $\mathbf{X}_j$, for $j = 1, \ldots, J$, conditionally on $(\mathbf{Y}_j, \mathbf{V}_j, \boldsymbol{\phi}_j, \alpha_j, \beta_j, \delta_j, \boldsymbol{\xi}_j, \mathbf{W}, \mathbf{S}_j)$.

The trajectory of the latent state variables $\mathbf{S}_j = (s_{j,1}, \ldots, s_{j,T})'$, for each j, is obtained using the FFBS algorithm (Frühwirth-Schnatter, 2006).

A.2 Sampling $\boldsymbol{\delta}_w$ and $\boldsymbol{\delta}_j$

The posterior distribution for δ_w and δ_j, for each $j = 1, \ldots, J$, are

$$P(\delta_w|\mathbf{W}, \alpha_w, \beta_w) \propto T\mathcal{G}a(\delta_w|a_{\delta_w}, b_{\delta_w}; \tau_w) \prod_{t=1}^{T} \text{NcGa}(w_t|\alpha_w, \beta_w w_{t-1}, \delta_w),$$

$$P(\delta_j|\mathbf{X}_j, \alpha_j, \beta_j) \propto T\mathcal{G}a(\delta_j|a_\delta, b_\delta; \tau) \prod_{t=1}^{T} \text{NcGa}(x_{j,t}|\alpha_j, \beta_j x_{j,t-1}, \delta_j).$$

We sample from these posterior distributions using an adaptive random walk Metropolis–Hastings (aRWMH) step with truncated lognormal proposal.

A.3 Sampling α_w and $\boldsymbol{\alpha}_j$

The posterior distributions for α_w and α_j, for each $j = 1, \ldots, J$, are

$$P(\alpha_w|\mathbf{W},\beta_w,\delta_w) \propto \mathcal{G}a(\alpha_w|a_{\alpha_w},b_{\alpha_w})\prod_{t=1}^{T}\text{NcGa}(w_t|\alpha_w,\beta_w w_{t-1},\delta_w),$$
$$P(\alpha_j|\mathbf{X}_j,\beta_j,\delta_j) \propto \mathcal{G}a(\alpha_j|a_\alpha,b_\alpha)\prod_{t=1}^{T}\text{NcGa}(x_{j,t}|\alpha_j,\beta_j x_{j,t-1},\delta_j).$$

We sample from them using an aRWMH step with lognormal proposal.

A.4 Sampling β_w and β_j

The posterior distributions for β_w and β_j, for each $j = 1,\ldots, J$, are

$$P(\beta_w|\mathbf{W},\alpha_w,\delta_w) \propto \mathcal{G}a(\beta_w|a_{\beta_w},b_{\beta_w})\prod_{t=1}^{T}\text{NcGa}(w_t|\alpha_w,\beta_w w_{t-1},\delta_w),$$
$$P(\beta_j|\mathbf{X}_j,\alpha_j,\delta_j) \propto \mathcal{G}a(\beta_j|a_\beta,b_\beta)\prod_{t=1}^{T}\text{NcGa}(x_{j,t}|\alpha_j,\beta_j x_{j,t-1},\delta_j).$$

We sample from them using an aRWMH step with lognormal proposal.

A.5 Sampling η_j and γ_j

The posterior distribution for η_j and γ_j, for each j, are given by

$$P(\eta_j\,|\,\xi_{j,2},\gamma_j) \propto \mathcal{G}a\left(a_\eta+\gamma_j(c_\gamma+1),\frac{b_\eta}{1+b_\eta\xi_{j,2}}\right), \quad P(\gamma_j\,|\,\xi_{j,2},\eta_j) \propto \frac{\left(a_\gamma \eta_j^{c_\gamma+1}\xi_{j,2}\right)^{\gamma_j}}{\Gamma(\gamma_j)^{b_\gamma+1}}$$

We sample from the latter distribution via the inverse transform method.

A.6 Sampling $\xi_{j,2}$

We sample from the posterior distribution of the jump sizes $\xi_{j,2}$, for each j, using an aRWMH step with Gamma proposal

$$P(\xi_{j,2}\,|\,\mathbf{Y}_j,\mathbf{W},\mathbf{X}_j,\eta_j,\gamma_j,\mathbf{S}_j) \propto \xi_{j,2}^{\gamma_j-1}\exp\left(-\xi_{j,2}\left(\eta_j+\sum_{t\in\mathcal{T}_{j,2}}x_{j,t}\right)\right)\prod_{t\in\mathcal{T}_{j,2}}(\overline{w}_t+x_{j,t}(1+\xi_{j,2}))^{y_{j,t}}$$

A.7 Sampling $\boldsymbol{\phi}_z$ and $\boldsymbol{\phi}_j$

The posterior distribution of the coefficients of the covariates, $\boldsymbol{\phi}_z$ and $\boldsymbol{\phi}_j$, for $j = 1, \ldots, J$, are given by

$$P(\boldsymbol{\phi}_z|\mathbf{Z}, \mathbf{V}_z, \mathbf{W}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\phi}_z' \underset{!!}{\Sigma}{}_z^{-1} \boldsymbol{\phi}_z - 2\boldsymbol{\phi}_z' \underset{!!}{\Sigma}{}_z^{-1} \underset{!!}{\boldsymbol{\phi}}_z)\right) \cdot \prod_{t=1}^{T} (w_t + \exp(\mathbf{v}_{z,t}' \boldsymbol{\phi}_z))^{z_t} \exp\left(-w_t - \exp(\mathbf{v}_{z,t}' \boldsymbol{\phi}_z)\right),$$

$$P(\boldsymbol{\phi}_j|\mathbf{Y}_j, \mathbf{V}_j, \mathbf{X}_j, \mathbf{W}, \boldsymbol{\xi}_j, \mathbf{S}_j) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\phi}_j' \underset{!!}{\Sigma}{}_\phi^{-1} \boldsymbol{\phi}_j - 2\boldsymbol{\phi}_j' \underset{!!}{\Sigma}{}_\phi^{-1} \underset{!!}{\boldsymbol{\phi}}_j)\right) \cdot \prod_{t=1}^{T} (\tilde{x}_{j,t} + \exp(\mathbf{v}_{j,t}' \boldsymbol{\phi}_j))^{y_{j,t}} \exp(-\tilde{x}_{j,t} - \exp(\mathbf{v}_{j,t}' \boldsymbol{\phi}_j)).$$

To improve the mixing of the chain, we use an aRWMH step with a Normal proposal to sample element wise from the posterior distribution of $\boldsymbol{\phi}_z$ and $\boldsymbol{\phi}_j$.

A.8 Posterior for Λ_j

The posterior distribution for each row l of each transition matrix Λ_j is

$$P(\boldsymbol{\lambda}_{j,l}|\mathbf{S}_j) \propto \mathcal{D}ir(\underset{!!}{\lambda}_{j1} + N_{l1}(\mathbf{S}_j), \ldots, \underset{!!}{\lambda}_{jL} + N_{lL}(\mathbf{S}_j)).$$