A Hierarchical Model of Nonhomogeneous Poisson Processes for Twitter Retweets

Article *in* Journal of the American Statistical Association · February 2018

DOI: 10.1080/01621459.2019.1585358

2 authors, including:

Clement Lee
Lancaster University
12 PUBLICATIONS **17** CITATIONS

SEE PROFILE

A hierarchical model of non-homogeneous Poisson processes for Twitter retweets

Clement Lee[1,2] and Darren J Wilkinson[1]

[1]School of Mathematics, Statistics and Physics, Newcastle University

[2]Open Lab, Newcastle University

February 7, 2018

Abstract

We present a hierarchical model of non-homogeneous Poisson processes (NHPP) for information diffusion on online social media, in particular Twitter retweets. The retweets of each original tweet are modelled by a NHPP, for which the intensity function is a product of time-decaying components and another component that depends on the follower count of the original tweet author. The latter allows us to explain or predict the ultimate retweet count by a network centrality-related covariate. The inference algorithm enables the Bayes factor to be computed, in order to facilitate model selection. Finally, the model is applied to the retweet data sets of two hashtags.

Keywords: Twitter; Hierarchical model; Non-homogeneous Poisson process; Model selection; Markov chain Monte Carlo

1 Introduction

Statistical modelling of online social media such as Twitter has become increasingly popular, because of the richness and availability of the data in temporal and topological aspects. A common approach to modelling temporal dynamics is the use of one-dimensional non-homogeneous Poisson Processes (NHPP), which are one prominent kind of stochastic process. Specifically, if a sequence of events is assumed to arise from a NHPP with intensity function $h(t) \geq 0$, the random variable of the number of events within

the interval $[t_1, t_2]$ will follow a Poisson distribution with mean $\int_{t_1}^{t_2} h(t)dt$, and is independent of the random number of events in any other disjoint interval. The special case where $h(t)$ is constant over time is called the homogeneous Poisson process (HPP).

Examples of using the HPP for Twitter data include Sakaki et al. (2010), Perera et al. (2010), Kumar et al. (2014, 2015), and Mahmud et al. (2013), but it usually does not describe data realistically because it assumes the interarrival times of events are independent and identically distributed (iid) exponential random variables. Sanli and Lambiotte (2015) observe bursty dynamics and temporal fluctuations in tweets with hashtag #`ledebat`, over the two-week period leading to the 2012 French presidential election, and show the departure of the data from one simulated from a HPP without fitting a stochastic process. It is therefore natural that the more general NHPP, or extensions thereof, is more often used in the literature. For example, in Smid et al. (2011), a NHPP is being used for semi-supervised detection of an anomaly in pollution-related tweets.

Quite often the intensity function $h(t)$ is specifically designed or chosen to capture temporal patterns observed in the data. For example, Mathiesen et al. (2013) and Mollgaard and Mathiesen (2015) both fit a NHPP to the occurrences of international brand names on Twitter, which exhibit strongly correlated used behaviour and bursty collective dynamics over time. The former incorporate long range temporal correlations in $h(t)$, resulting in interarrival times that are marginally distributed according to the power law, while the latter consider $h(t)$ as a product of stochastic global user interest and approximately deterministic user activity over time. Such a way of splitting $h(t)$ into two components is also seen in Shen et al. (2014), and Mathews et al. (2017). The former fit a NHPP to the popularity dynamics of Twitter hashtags and Physical Review papers, where $h(t)$ is a product of a decreasing function of time and a term increasing linearly with the number of events, intended to capture the effect of how the attractiveness of an individual item ages, and the effect of preferential attachment, respectively. The latter fit a NHPP to the retweets of popular Twitter users with $h(t)$ proportional to the product of $t^{-\lambda}$ and $e^{-\theta t}$, and attempt to explain the two components by a decision-based queueing process, rather than preferential attachment, and loss of interest over time, respectively. The model by Mollgaard and Mathiesen (2015) can also be viewed as extending the NHPP by incorporating stochasticity in $h(t)$, which is also seen in some other models. Pozdnoukhov and Kaiser (2011) use a Markov-modulated Poisson process, in which $h(t)$ varies according to a Markov process, in their application of identification and spatio-temporal analysis of topics on Twitter. Bao et al. (2015) propose a self-excited Hawkes process (SEHP), in which $h(t)$ jumps simultaneously when an event occurs and decays before the next event occurs, and argue that such their model outperforms the one by Shen et al. (2014), in terms of prediction accuracy, when applied to the same set of data.

Twitter data usually comes with information such as number of followers or even who follows whom, that is, the directed edges in the user network, therefore enabling modelling of its social network, static or dynamic. For example, Bhamidi et al. (2015) collect tweets of specific topics associated with competing hashtags, and observe the departure of the degree distribution of the retweet network from one predicted by the classical preferential attachment model (Barabási and Albert, 1999). They propose a variant called the Superstar model, in which a vertex enters the network by connecting to either the lone superstar, with the same probability across all vertices, or the rest of the network otherwise, according to original preferential attachment rule.

Whenever the data permits, it is natural to extend a model to account for the temporal dynamics and the network structure simultaneously, see, for example, Li et al. (2014). Xie et al. (2011) and Wu et al. (2011) build a framework for data on Sina weibo, a Chinese counterpart of Twitter, that divides users into communities and models information generation, receiving, and processing and diffusion by the power law, a HPP, and a multiplicative model of individual reading habits and relation strength, respectively. It is also possible to model network structure and information diffusion simultaneously without using time as a dimension. Both Li et al. (2012) and Nishi et al. (2016) use a Galton-Watson branching process model, for data of video contents shared on online social networks, and reply trees in Twitter, respectively.

Usually and implicitly assumed in the models aforementioned is that the network, if concerned, remains unchanged throughout the observation period, which may be unrealistic for Twitter data given the ease of following other users. Therefore efforts have been made to model the dynamics of information diffusion and network evolution simultaneously, as it is natural to conjecture that they co-evolve over time. Antoniades and Dovrolis (2015) propose a tweet-retweet-follow model, which is characterised by events of a follower of a retweeter becoming also a follower of the original tweet author, conditional on the original tweet being created and retweeted. Farajtabar et al. (2015) consider the follower and the retweet adjacency matrices, and model the co-evolution through a system of dynamic equations of these two matrices. Srijith et al. (2017) assume no network evolution but incorporate the influence between users according to the network in a multivariate Hawkes process model, in which each user-topic pair has its own intensity function. Lim et al. (2016) introduce a Twitter-Network topic model, which comprises a hierarchical Poisson-Dirichlet process model for the text and hashtags, and a Gaussian process based random function model for the followers network, and is applied to a data set of tweets with certain keywords.

Instead of modelling temporal and network dynamics merely according to some stochastic processes, one can look into how network summary measures, such as number of followers, and other variables affect either or both of them, thus identifying useful covariates for predictions for retweet behaviour and network

influence. Sutton et al. (2014) employ a negative binomial regression model for the retweet count, to investigate how message content/style and public attention to tweets relate to the retweet activity in a disaster. Zhu et al. (2011) apply a logistic regression model to the data of whether a tweet is being retweeted from the point of view of a *follower*. Hong et al. (2013) include a regression part in their co-factorization machines, which are for discovering topics users are interested in. They suggest that both network measures and content are important in determining retweets. However, the relationships among the covariates are not reported in all three analyses, thus presenting the risk of potential collinearity and overfitting.

Commonly observed and modelled in the aforementioned literature is the power law phenomenon. Examples in temporal aspects include interarrival times (Götz et al., 2009, Xie et al., 2011, Wu et al., 2011), and tweet or retweet rate (Mathiesen et al., 2013, Mathews et al., 2017), while examples in topological aspects include network degree (Li et al., 2014) and size and depth of reply trees (Nishi et al., 2016). Regarding network influence, the power law has been observed in retweet count (Hong et al., 2013), count of in-links for blogs (Götz et al., 2009), view count of videos (Miotto et al., 2017), and citation count (Shen et al., 2014). Interestingly, there have been no studies on the relationships among these variables following the power law.

The research reported in this article stemmed from investigating all tweets (both original and retweets) with two specific hashtags. Compared to the analysis by Shen et al. (2014), we dig one level deeper as we model the original tweets, simply called the originals hereafter, and retweets by two separate stochastic processes. Common between the two levels of modelling is the use of the hybrid process, a particular NHPP with $h(t)$ adopted from, for example, Mathews et al. (2017), but without resorting to discretising the data when it comes to inference. While it is straightforward to fit a hybrid process to the originals, as we will illustrate in Section 2, the novelty of this article is the hierarchical modelling of retweets. Specifically, all retweets of each original are modelled by a NHPP, with a latent component in $h(t)$ that depends on the follower count (of the author of the original) and determines the ultimate retweet count. All the retweet processes are in turn enveloped in one single hierarchical model so that information can be pooled to estimate the parameters.

There are a few merits of including follower count, which is essentially the in-degree of a user, and retweet count in the way described above, both of which are observed to follow the power law empirically. First, it presents a network centrality-related covariate as the potential driving force of retweet behaviour or network influence, while simultaneously modelling the temporal dynamics. Second, such a way of incorporating network summary measures enables us to capture any effect attributed to the power law phenomenon, while avoiding the overhead of an explicit network structure, which is usually computa-

tionally expensive to construct. Furthermore, directly using the follower count, which can vary over the observation period even for the same user, already partially accounts for the effect of network evolution over time. Finally, this model is generative in the sense that, conditional on the follower count of the authors of the originals (which can be easily generated by the power law), we can simulate a realistic process of processes, each of which corresponds to how retweets of a particular original grow over time.

The rest of the article is divided as follows. The hybrid process and its special case are introduced in Section 2, with applications to data of originals. In Section 3 we use a censored regression model to explain retweet count by follower count. The hierarchical model primarily for retweets is introduced in Section 4, with its likelihood derived. The inference algorithm is outlined in Section 5, together with Bayes factor calculations for model selection. The model is applied to two real data sets of retweets in Section 6. Section 7 concludes the article.

2 The hybrid and power law processes

Consider a NHPP with $h(t) = \gamma t^{-\lambda} e^{-\theta t}$, where $\gamma > 0, \theta \geq 0$ and $\lambda < 1$, which is called the hybrid process hereafter. It is equivalent to the "power law with exponential cutoff" function by Mathews et al. (2017), but is different from the doubly stochastic processes introduced in Section 1 as $h(t)$ is deterministic. The cumulative intensity is given by

$$H(t) := \int_0^t h(u)du = \begin{cases} \gamma\Gamma(1-\lambda, \theta t)\theta^{\lambda-1}, & \theta > 0, \\ \gamma t^{1-\lambda}/(1-\lambda), & \theta = 0, \end{cases} \tag{1}$$

where $\Gamma(x, y)$ is the lower incomplete Gamma function such that $\lim_{y\to\infty} \Gamma(x, y)$ is equal to the Gamma function $\Gamma(x)$. Now assume a sequence of n events is generated from the hybrid process in the time interval $[0, T]$, in which the i-th event occurs at time $t_i (i = 1, 2, \ldots, n)$, so that $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \leq T$. It is straightforward to write down the likelihood function:

$$\begin{aligned} f(t_1, t_2, \ldots, t_n | \lambda, \theta, \gamma) &:= \exp\left[-H(T)\right] \times \prod_{i=1}^{n} h(t_i) \\ &= \begin{cases} \exp\left[-\gamma\Gamma(1-\lambda, \theta T)\theta^{\lambda-1}\right] \times \prod_{i=1}^{n} \gamma t_i^{-\lambda} e^{-\theta t_i}, & \theta > 0, \\ \exp\left[-\gamma T^{1-\lambda}/(1-\lambda)\right] \times \prod_{i=1}^{n} \gamma t_i^{-\lambda}, & \theta = 0. \end{cases} \end{aligned} \tag{2}$$

When $\theta = 0$, the hybrid process becomes the power law process (Bar-Lev et al., 1992). Each of the interarrival times follow a truncated Weibull distribution (Yakovlev et al., 2005). The power law process is different from a renewal process with power law distributed interarrival times, such as the event-modulated Poisson process termed by Masuda and Rocha (2017).

Going back to the aforementioned sequence of events, if we want to check if it is generated by the power law process, we can fit the hybrid process and then formally test whether θ is 0 using whatever estimation approach. However, there is also a diagnostic plot for checking whether the power law process is appropriate with no model fitting required. Observe that the expected number of events at t_i $(i = 1, 2, \ldots)$, denoted by $\mathbb{E}(N(t_i))$, should be close to i, where $N(t)$ is the number of events in the interval $[0, t]$. Under the power law process, the former is given by $\mathbb{E}(N(t_i)) = H(t_i) = \dfrac{\gamma t_i^{1-\lambda}}{1-\lambda}$, and so we have the following approximations:

$$\frac{\gamma t_i^{1-\lambda}}{1-\lambda} \approx i \quad \Leftrightarrow \quad \frac{t_i}{i} \approx \frac{1-\lambda}{\gamma} t_i^{\lambda} \quad \Leftrightarrow \quad \log\left(\frac{t_i}{i}\right) \approx \log\left(\frac{1-\lambda}{\gamma}\right) + \lambda \log t_i. \tag{3}$$

This means that plotting t_i/i, which is termed mean time between failures (MTBF) in reliability theory, against t_i on the log-log scale should give approximately a straight line with slope λ and intercept $\log(1-\lambda) - \log(\gamma)$. This is called the Duane plot (Duane, 1964), which serves as a useful tool for diagnosing if the power law process describes the data well, and is similar to judging whether the Weibull distribution is useful according to the survival log-log plot in survival analysis.

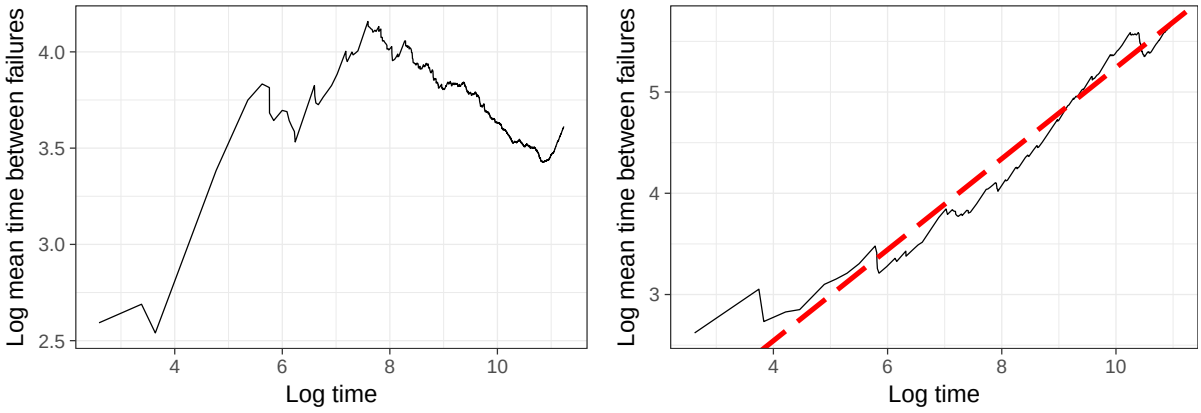

Figure 1: Duane plots of originals (left), where time is relative to the start of observation period, and retweets of the top original (right), where time is relative to when the original is tweeted. Overlaid is the theoretical line (dashed) according to (3) with $(\lambda, \gamma) = (\hat{\lambda}, \hat{\gamma})$.

A set of tweets with the hashtag #`thehandmaidstale` was collected on 2017-06-14 for 21 hours after one episode of the relevant TV series was broadcasted. The Duane plot for the 2043 originals is shown on the left of Figure 1, where linearity is only observed at certain intervals. We also formally fit the power law process to the data, by maximising the (log-)likelihood in the second line of (2) with respect to λ and γ simultaneously, yielding $(\hat{\lambda}, \hat{\gamma}) = (0.024, 0.034)$, where $\hat{\eta}$ denotes the maximum likelihood estimate (MLE) for any parameter η, and a maximised log-likelihood of -9422.866. Fitting the hybrid process instead gives $(\hat{\lambda}, \hat{\theta}, \hat{\gamma}) = (-0.625, 3.152 \times 10^{-5}, 1.418 \times 10^{-4})$ and a maximised log-likelihood of -9336.362. While the difference in the maximised log-likelihood between the power law process and the hybrid process suggests that the former is inadequate, the latter does not necessarily describe the data well enough.

On the right of Figure 1 is the Duane plot for the retweets of the top original (with the highest retweet count of 204), which shows linearity over the whole observation period apart from a few small troughs and a seemingly increasing positive slope, suggesting that the power law process may be sufficient compared to the more general hybrid process. This is confirmed by fitting the latter to the data, which gives $(\hat{\lambda}, \hat{\theta}, \hat{\gamma}) = (0.45, 0, 0.261)$, and no reduction in the maximised log-likelihood compared to the respective power law process fit. The dashed line overlaying the Duane plot represents (3) with $(\lambda, \gamma) = (\hat{\lambda}, \hat{\gamma})$, and its proximity provides further support to the adequacy of the power law process.

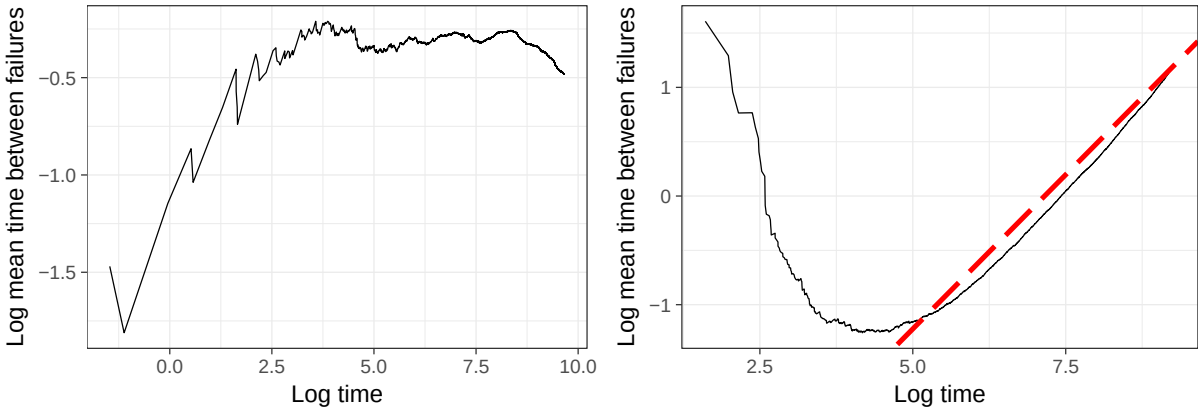

Figure 2: Duane plots of originals (left) and retweets of the top original (right) for the #gots7 data, overlaid by the theoretical line (dashed) according to (3).

To examine potentially different tweeting behaviour of different hashtags, a set of tweets with the hashtag #`gots7` was collected on 2017-07-16 for around 4.4 hours *before* the 7$^{\text{th}}$ season premiere of the TV series Game of Thrones was broadcasted. For the 25420 originals, fitting the power law process gives $(\hat{\lambda}, \hat{\gamma}) = (-0.131, 0.514)$, while the hybrid process does not improve the fit with the same point estimates and $\hat{\theta} = 0$. For the 3204 retweets of the top original, fitting the power law process gives maximised log-likelihood -5545.778 and $(\hat{\lambda}, \hat{\gamma}) = (0.568, 25.104)$, which are used to obtain the theoretical line overlaid in the Duane plot in Figure 2. The slight concavity of the Duane plot in the overlapping interval indicates possible inadequacy of the power law process and potential improvement by the hybrid process, which is supported by fitting the latter obtain maximised log-likelihood -5401.13 and $(\hat{\lambda}, \hat{\theta}, \hat{\gamma}) = (0.408, 1.562 \times 10^{-4}, 12.932)$.

For each of the two hashtags considered, whether the power law process is adequate for the retweets of the top original should not be assumed to automatically apply to the retweets of every other original. For exploratory purposes, we overlay the Duane plots of retweets of the top 13 originals in Figure 3, all with over 300 retweets. That the slope is more similar across different Duane plots than the position is suggests that respective fits by the power law process (or the hybrid process) will give more similar estimates of λ than of γ. In terms of actual modelling, we will stick to the hybrid process in the hierarchical model for the retweets introduced in the next section, and formally test whether θ, which will be universal to all originals, is equal to 0 in Section 5 through model selection.

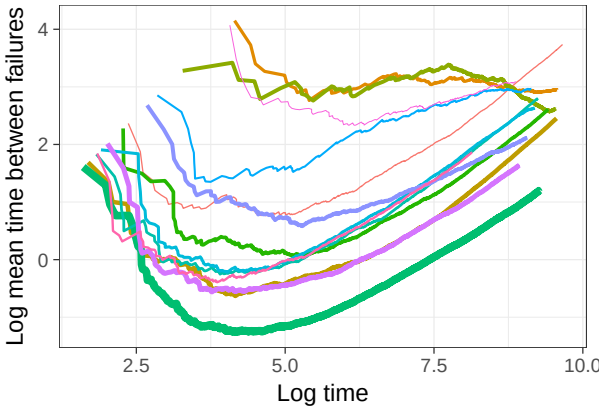

Figure 3: Duane plots of retweets of the top 13 originals for the #gots7 data. The thicker the line is, the more retweets the original has.

3 Censored regression model

In this section we directly explain the retweet count, which is the outcome of the process generating retweets, by the follower count, which is observed once the original is tweeted. For the i-th original, we assume that m_i retweets are observed in the interval $[t_i, T]$, where the j-th retweet occurs at time $t_{ij} (j = 1, 2, \ldots, m_i)$, so that $t_i \leq t_{i1} \leq t_{i2} \leq \cdots \leq t_{im_i} \leq T$. We also assume that the author of this i-th original has $x_i \geq 0$ followers at time t_i, and define $m_i^* = \log(1 + m_i)$ and $x_i^{'} = \log(1 + x_i)$ to be the "transformed" retweet count and follower count, respectively. Finally, we define $x_i^* = x_i^{'} - \frac{1}{n}\sum_{k=1}^{n} x_k^{'}$ to be the "mean-centred" follower count. The scatterplots of m_i^* against x_i^* for the retweets of the aforementioned data sets are shown in Figure 4.

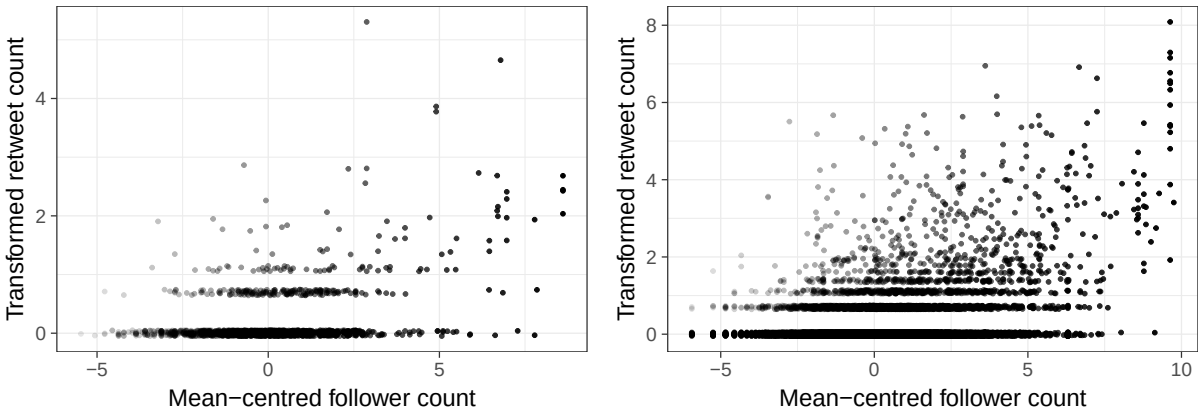

Figure 4: Transformed retweet count against mean-centred follower count for #thehandmaidstale data (left) and #gots7 data (right).

To model such relationship, due to the positivity of the response, a better alternative to a linear regression model is the censored regression model (Tobin, 1958):

$$m_i^* = \max\left(0, \alpha + \beta x_i^* + \epsilon_i\right), \tag{4}$$

where $$\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \tau^{-1}), \qquad i = 1, 2, \ldots, n, \tag{5}$$

where α, β and τ are parameters and ϵ_i is a random error. That x_i^* is used in preference to $x_i^{'}$ is due to mean-centring that orthogonalises α and β in inference. As with the linear regression model, the censored regression model can include additional terms with higher powers of x_i^* to potentially improve its fit. For the #thehandmaidstale data, fitting the censored regression with a linear, quadratic and cubic polynomial of x_i^* results in a maximum log-likelihood of 828.398, 841.473 and 843.761, respectively. For the #gots7 data, the linear, quadratic and cubic fits gives a maximum log-likelihood of 10236.327, 10439.882 and 10440.032, respectively. As the quadratic fit outperforms the linear fit substantially and is bettered by the cubic fit slightly for both data sets, its model structure

$$m_i^* = \max\left(0, \alpha + \beta x_i^* + \kappa (x_i^*)^2 + \epsilon_i\right), \tag{6}$$

where κ is the additional parameter for the quadratic term, will be incorporated in the full hierarchical model of the process of retweets, with a slight modification, in the following section.

4 Hierarchical model and likelihood

For convenience, the terminology and notation in Sections 2 and 3 are retained, but no model is assumed for how the *originals* are generated as it is not the concern of this section. Instead, the *retweets* of the i-th original are assumed to arise from a hybrid process with intensity

$$h_i(t) = \begin{cases} (e^{\delta_i} - 1)(t - t_i)^{-\lambda} e^{-\theta(t - t_i)}, & t \geq t_i, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The shift from t to $t - t_i$ in $h_i(t)$ is due to the process of retweets taking place relative to the time original i is tweeted. While λ and θ are universal to the process of each original, δ_i is dependent on x_i^* and given by

$$\delta_i = \max\left(0, \alpha + \beta x_i^* + \kappa \left(x_i^*\right)^2 + \epsilon_i\right), \tag{8}$$

which means $$\alpha + \beta x_i^* + \kappa \left(x_i^*\right)^2 + \epsilon_i > 0 \quad \Leftrightarrow \quad \delta_i > 0 \quad \Leftrightarrow \quad e^{\delta_i} - 1 > 0 \quad \Leftrightarrow \quad h_i(t) > 0, \tag{9}$$

and $$\alpha + \beta x_i^* + \kappa \left(x_i^*\right)^2 + \epsilon_i \leq 0 \quad \Leftrightarrow \quad \delta_i = 0 \quad \Leftrightarrow \quad e^{\delta_i} - 1 = 0 \quad \Leftrightarrow \quad h_i(t) = 0, \tag{10}$$

for $t \geq t_i$. This establishes a direct relationship between the linear predictor and the intensity, and (9) and (10) will be useful notation later. In situations where δ_i needs to be illustrated as a function of the parameters, ϵ_i and x_i^*, we will write $\delta_i(\alpha, \beta, \kappa, \epsilon_i, x_i^*)$ instead. Upon rearranging the cumulative density

at infinity, we have (for the case $\theta > 0$)

$$H_i(\infty) = (e^{\delta_i} - 1)\Gamma(1-\lambda)\theta^{\lambda-1},$$
$$\log\left[1 + \frac{H_i(\infty)}{\Gamma(1-\lambda)\theta^{\lambda-1}}\right] = \delta_i = \max\left(0, \alpha + \beta x_i^* + \kappa\left(x_i^*\right)^2 + \epsilon_i\right). \tag{11}$$

The corresponding equation for $\theta = 0$ can be derived similarly and is omitted here. If T is large enough, due to the decay of $h(t)$ over time, $H_i(\infty)$ should be close to $H_i(T - t_i)$, which should in turn be close to the retweet count m_i, if the hybrid process is the true underlying process. This gives the following approximation:

$$\log\left[1 + \frac{m_i}{\Gamma(1-\lambda)\theta^{\lambda-1}}\right] \approx \max\left(0, \alpha + \beta x_i^* + \kappa\left(x_i^*\right)^2 + \epsilon_i\right). \tag{12}$$

Comparing (12) with (6) shows a subtle difference between the original censored regression model and our model, as ours links the transformed *and rescaled* retweet count with the mean-centred follower count. While it is possible to include the normalising constant $\Gamma(1-\lambda)\theta^{1-\lambda}$ in (7) in order to align the two models and allow a natural interpretation of $e^{\delta_i} - 1$ as the ultimate retweet count, doing so will not yield the power law process when $\theta \to 0$, therefore justifying its exclusion in the intensity function.

Before expressing the likelihood as a function of the parameters and the random errors $\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, which are treated as latent variables, we first define $\boldsymbol{m} := (m_1, m_2, \ldots, m_n)$, $\boldsymbol{t} := \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$, where $\boldsymbol{t}_i := (t_i, t_{i1}, t_{i2}, \ldots, t_{im_i}), \boldsymbol{x}^* := (x_1^*, x_2^*, \ldots, x_n^*), \boldsymbol{\eta}_0 := (\alpha, \beta, \kappa, \lambda, \tau, \boldsymbol{\epsilon})$, and $\boldsymbol{\eta}_1 := (\alpha, \beta, \kappa, \lambda, \tau, \boldsymbol{\epsilon}, \theta)$. Furthermore, we introduce a parameter M, which can take value 0 or 1, to represent model *choice*. When $M = 0$, the hierarchical model of the power law process is the true model with parameter vector $\boldsymbol{\eta}_0$, in which θ is set to 0 and removed. When $M = 1$, the hierarchical model of the hybrid process with $\theta > 0$ is the true model with parameter vector $\boldsymbol{\eta}_1$. By treating the nested models as two competing models, the problem of testing whether $\theta = 0$ becomes a problem of model selection, which can be achieved by utilising the output of the inference algorithm outlined in Section 5. Our algorithm requires the likelihood for each model as a function of $\boldsymbol{\eta}_M$:

$$\begin{aligned}
&f(\boldsymbol{m}, \boldsymbol{t} | \boldsymbol{x}^*, \boldsymbol{\eta}_0, M = 0) \\
&\quad = \exp\left(-(1-\lambda)^{-1} \sum_{i=1}^{n} \left[\left(e^{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)} - 1\right)(T - t_i)^{1-\lambda}\right]\right) \\
&\qquad \times \prod_{i:m_i>0} \left[\mathbf{1}_{\{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)>0\}} \left(e^{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)} - 1\right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda}\right],
\end{aligned} \tag{13}$$

$$\begin{aligned}
&f(\boldsymbol{m}, \boldsymbol{t} | \boldsymbol{x}^*, \boldsymbol{\eta}_1, M = 1) \\
&\quad = \exp\left(-\theta^{\lambda-1} \sum_{i=1}^{n} \left[\left(e^{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)} - 1\right)\Gamma\left(1-\lambda, \theta(T-t_i)\right)\right] - \theta \sum_{i:m_i>0} \sum_{j=1}^{m_i} (t_{ij} - t_i)\right)
\end{aligned}$$

$$\times \prod_{i:m_i>0} \left[\mathbf{1}_{\{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)>0\}} \left(e^{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)} - 1 \right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda} \right], \tag{14}$$

where $\mathbf{1}_{\{A\}}$ is the indicator function of event A. The derivations of (13) and (14) are detailed in Appendix A. Note that, even though τ is seen in neither (13) nor (14) because of independence between τ and the data conditional on $\boldsymbol{\epsilon}$, it is included in $\boldsymbol{\eta}_M$ for notational convenience.

5 Inference and the Bayes factor

The presence of the latent variables $\boldsymbol{\epsilon}$ and the problem of model selection between $M = 0$ and $M = 1$ prompt us to consider Bayesian inference for the proposed hierarchical model. We first assign the following independent and vaguely informative priors:

$$\begin{aligned} \alpha &\sim \mathrm{N}\left(\mu_\alpha = 0, \tau_\alpha^{-1} = 10^4\right), \\ \beta &\sim \mathrm{N}\left(\mu_\beta = 0, \tau_\beta^{-1} = 10^4\right), \\ \kappa &\sim \mathrm{N}\left(\mu_\kappa = 0, \tau_\kappa^{-1} = 10^4\right), \\ (1-\lambda) &\sim \mathrm{Gamma}\left(a_\lambda = 1, b_\lambda = 0.001\right), \\ \tau &\sim \mathrm{Gamma}\left(a_\tau = 1, b_\tau = 0.001\right), \end{aligned} \tag{15}$$

where τ_X^{-1} is the variance of a random variable $X \sim \mathrm{N}\left(\mu_X, \tau_X^{-1}\right)$, and a_Y/b_Y is the mean of a random variable $Y \sim \mathrm{Gamma}\left(a_Y, b_Y\right)$. Also, under $M = 1$, we assign a Gamma$(a_\theta = 1, b_\theta = 0.001)$ prior for θ. As the parameter space is not the same for $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$, we denote $\boldsymbol{\eta}_{\setminus M}$ as the subset of $\boldsymbol{\eta}_M$ not in $\boldsymbol{\eta}_{1-M}$, which means $\boldsymbol{\eta}_{\setminus 0} = \theta$ and $\boldsymbol{\eta}_{\setminus 1} = \{\}$, the null set. Assuming conditional independence of $\boldsymbol{\eta}_M$ and $\boldsymbol{\eta}_{\setminus M}$ given M, the joint posterior of $\boldsymbol{\eta}_M$, $\boldsymbol{\eta}_{\setminus M}$ and M is

$$\begin{aligned} \pi(\boldsymbol{\eta}_M, \boldsymbol{\eta}_{\setminus M}, M | \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}^*) &\propto \pi(\boldsymbol{m}, \boldsymbol{t}, \boldsymbol{\eta}_M, \boldsymbol{\eta}_{\setminus M}, M | \boldsymbol{x}^*) \\ &= f(\boldsymbol{m}, \boldsymbol{t} | \boldsymbol{x}^*, \boldsymbol{\eta}_M, M) \times \pi(\boldsymbol{\eta}_M | M) \times \pi(\boldsymbol{\eta}_{\setminus M} | M) \times \pi(M). \end{aligned} \tag{16}$$

The "true" prior of $\boldsymbol{\eta}_M$ under M is given by

$$\pi(\boldsymbol{\eta}_M | M) = \begin{cases} \pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}|\tau) \times \pi_\alpha(\alpha)\pi_\beta(\beta)\pi_\kappa(\kappa)\pi_\lambda(\lambda)\pi_\tau(\tau), & M = 0, \\ \pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}|\tau) \times \pi_\alpha(\alpha)\pi_\beta(\beta)\pi_\kappa(\kappa)\pi_\lambda(\lambda)\pi_\tau(\tau)\pi_\theta(\theta), & M = 1, \end{cases} \tag{17}$$

where $\pi(\boldsymbol{\epsilon}|\tau)$ and the rest are given by (5) and (15), respectively, while the last component $\pi(M)$ of (16) is the prior probability that model M is true. For $M = 1$, the pseudoprior $\pi(\boldsymbol{\eta}_{\setminus M}|M)$ vanishes as $\boldsymbol{\eta}_{\setminus 1} = \{\}$, while for $M = 0$, the pseudoprior can be written as $\pi(\theta|M = 0)$ equivalently. We proceed to draw samples of $(\boldsymbol{\eta}_M, \boldsymbol{\eta}_{\setminus M}, M)$ using Markov chain Monte Carlo (MCMC), in which model selection is

facilitated by the modified version (Dellaportas et al., 2002) of Gibbs variable selection (Carlin and Chib, 1995). The MCMC algorithm is outlined as follows:

1. The current values in the chain are $\boldsymbol{\eta}_M$, $\boldsymbol{\eta}_{\setminus M}$ and M.
2. Draw $\boldsymbol{\eta}_M$ from its conditional posterior, with density proportional to $f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, \boldsymbol{\eta}_M, M)\times\pi(\boldsymbol{\eta}_M|M)$, by a fairly standard component-wise Metropolis-within-Gibbs (MWG) algorithm, the details of which are given in Appendix B. Denote the value by $\boldsymbol{\eta}_M^{'}$.
3. If $M = 0$, draw θ from its pseudoprior $\pi(\theta|M = 0)$. Denote the value by $\theta^{'}$, and write $\boldsymbol{\eta}_1^{'} = (\boldsymbol{\eta}_0^{'}, \theta^{'})$. If $M = 1$, write $\boldsymbol{\eta}_0^{'} = \boldsymbol{\eta}_{1,-\theta}^{'}$, that is, the proposed value of $\boldsymbol{\eta}_1$ with that of θ dropped, so that $\boldsymbol{\eta}_1^{'} = (\boldsymbol{\eta}_0^{'}, \theta^{'})$ still holds.
4. Draw M from its conditional posterior distribution $\pi(M|\boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}^*, \boldsymbol{\eta}_M, \boldsymbol{\eta}_{\setminus M})$. Essentially, set M to 0 and 1 with probabilities $\dfrac{A_0}{A_0 + A_1}$ and $\dfrac{A_1}{A_0 + A_1}$, respectively, where, using (16),

$$A_0 = f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, \boldsymbol{\eta}_0^{'}, M = 0)\ \pi(\theta^{'}|M = 0)\ \pi(M = 0),$$
$$A_1 = f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, \boldsymbol{\eta}_1^{'}, M = 1)\ \pi(\theta^{'}|M = 1)\ \pi(M = 1).$$

5. Denote the drawn value in step 4 by $M^{'}$. The current values are now $\boldsymbol{\eta}_M^{'}$, $\boldsymbol{\eta}_{\setminus M}^{'}$ and $M^{'}$.

The pseudoprior $\pi(\theta|M = 0)$ is chosen to be close to $\pi(\theta|M = 1, \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}^*)$, that is, the posterior under the competing model, for the sake of optimisation (Dellaportas et al., 2002, Carlin and Chib, 1995), which can be informed by a pilot run of the MWG algorithm in Appendix B for model 1 individually. As the priors for the overlapping parameters are the same for both models, only the pseudoprior $\pi(\theta|M = 0)$ and the prior $\pi(\theta|M = 1)$, the latter of which is the same as $\pi_\theta(\theta)$ in (17), are involved in step 4.

The draws of $\boldsymbol{\eta}_0$ in the above algorithm where $M = 0$ marginally represent an approximate sample from $\pi(\boldsymbol{\eta}_0|M = 0, \boldsymbol{m}, \boldsymbol{t})$, that is, its posterior distribution *under that model 0 is true*; likewise for $\boldsymbol{\eta}_1$. What is more important, however, is the empirical proportion of M, denoted by $\hat{\pi}(M|\boldsymbol{m}, \boldsymbol{t})$, as it approximates the posterior probability that model M is true. Finally, the Bayes factor is the ratio of the posterior odds to the prior odds:

$$B_{10} = \frac{\hat{\pi}(M = 1|\boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}^*)}{\hat{\pi}(M = 0|\boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}^*)} \Big/ \frac{\pi(M = 1)}{\pi(M = 0)}\ . \tag{18}$$

An alternative to Gibbs variable selection (GVS) for model selection is reversible jump Markov chain Monte Carlo (RJMCMC) (Green, 1995), which should theoretically give the same posterior probabilities for the model choice. It will be used to verify with the results of GVS in the application, and the details of its algorithm are given in Appendix C.

6 Application

Both the model-specific algorithms in Appendix B and the model selection algorithm in 5 are applied to the two data sets with different hashtags. For the #thehandmaidstale data, each of the three algorithms is applied to the times of creation of the 2043 originals, 265 of which have been retweeted at least once, and of their associated retweets, to obtain a single chain of 20000 iterations, upon thinning of 2000, after discarding the first 1000000 as burn-in. The individual model fits are reported in the form of traceplots and posterior densities of the parameters in Figure 5. While the inclusion of θ in model 1 makes a substantial difference in terms of the posterior densities of the other parameters, what is more important is how the evidence of each model weighs against each other. In the model selection algorithm, the prior probabilities $\pi(M = 0)$ and $\pi(M = 1)$ are chosen to be 0.95 and 0.05, respectively, to ensure sufficient mixing between the two states of M in the chain. Model 0 is selected for 14264 times out of 20000 iterations, meaning that B_{10} in (18) is estimated to be $\frac{5736}{14264} \Big/ \frac{0.05}{0.95} = 7.64$. The RJMCMC algorithm gives a similar estimate of $B_{10} = \frac{5940}{14060} \Big/ \frac{0.05}{0.95} = 8.027$. So, for the #thehandmaidstale data set, which consists of tweets for over 21 hours, the hybrid process hierarchical model is more appropriate. Such findings are consistent with the exponential cutoff phenomenon shown by Mathews et al. (2017) for tweets collected over a similar duration of 24 hours.

The #gots7 data set consists of 25420 originals, 3145 of which have been retweeted once, and 29751 retweets. To facilitate the much greater computational burden, for each of the two model-specific algorithms and the model selection algorithm, 40 chains are obtained, each of which contains 500 iterations, upon thinning of 1000, after discarding the first 500000 as burn-in. The traceplots and the posterior densities for the parameters are plotted in Figure 6. The proximity of the posterior densities for all parameters other than θ suggests that the inclusion of θ is less influential to the model fit for this data set than for the #thehandmaidstale data. This is supported by the model selection results via GVS. With the prior probabilities $\pi(M = 0)$ and $\pi(M = 1)$ chosen to be 10^{-9} and $1 - 10^{-9}$, respectively, model 0 is selected for 8983 times out of 20000 iterations in total, meaning that the Bayes factor is estimated to be $B_{10} = \frac{11017}{8983} \Big/ \left(\frac{1-10^{-9}}{10^{-9}}\right) = 1.226 \times 10^{-9}$. The RJMCMC algorithm gives a similar estimate of $B_{10} = \frac{10917}{9083} \Big/ \left(\frac{1-10^{-9}}{10^{-9}}\right) = 1.202 \times 10^{-9}$, meaning that model 0 is highly favoured. That the power law process hierarchical model is more appropriate for the #gots7 data set, which consists of tweets for about 4.4 hours, is also consistent with the findings by Mathews et al. (2017) for tweets collected over a similar duration of 3 hours.

To examine the goodness-of-fit of our model, for each data set, the retweet count estimated by the respective model selected by Gibbs variable selection, is plotted, with 95% credible intervals, against the actual

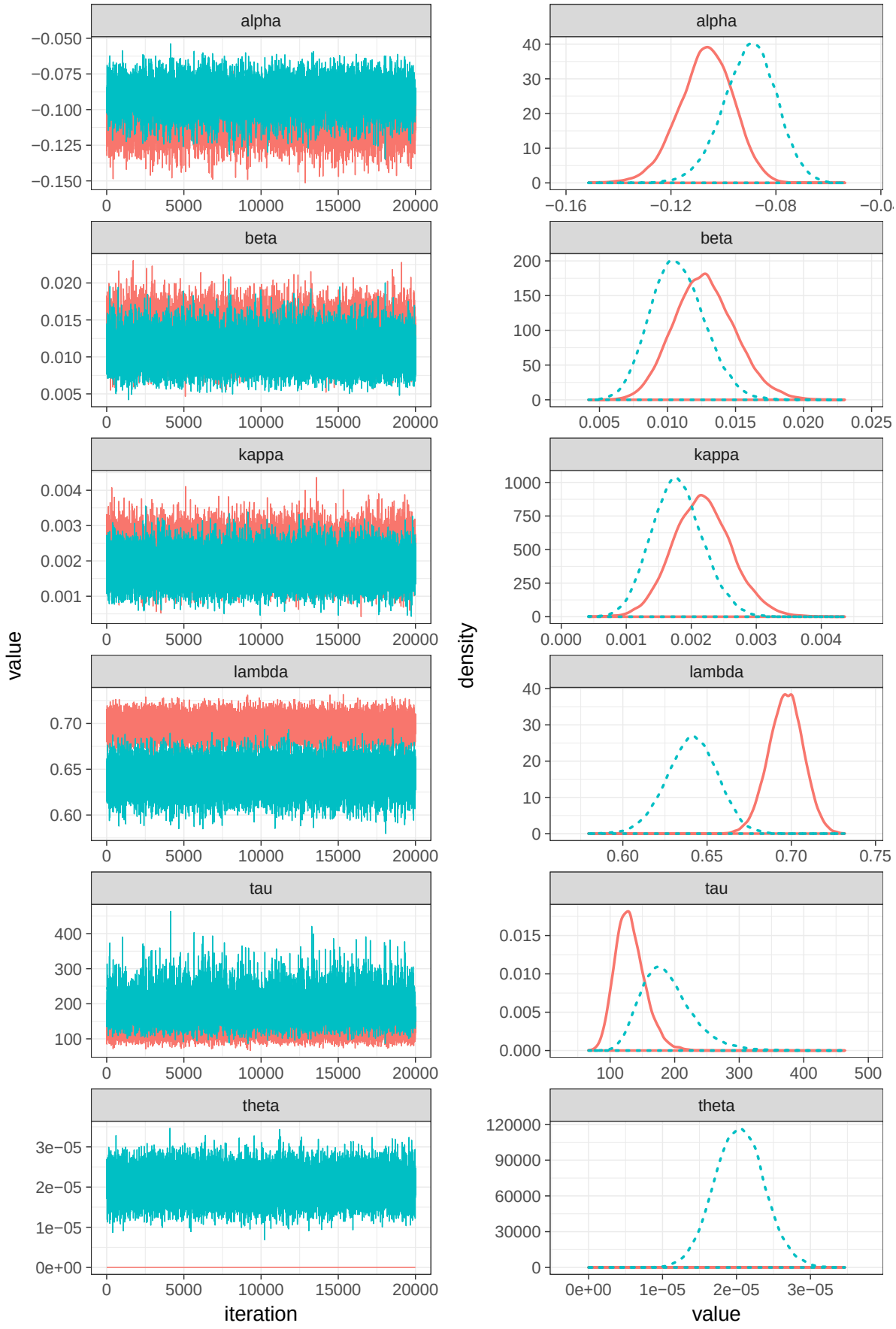

Figure 5: Traceplots (left) and posterior densities (right) of the parameters of models 0 (salmon, solid) and 1 (turquoise, dashed on the right) fitted to #thehandmaidstale data.

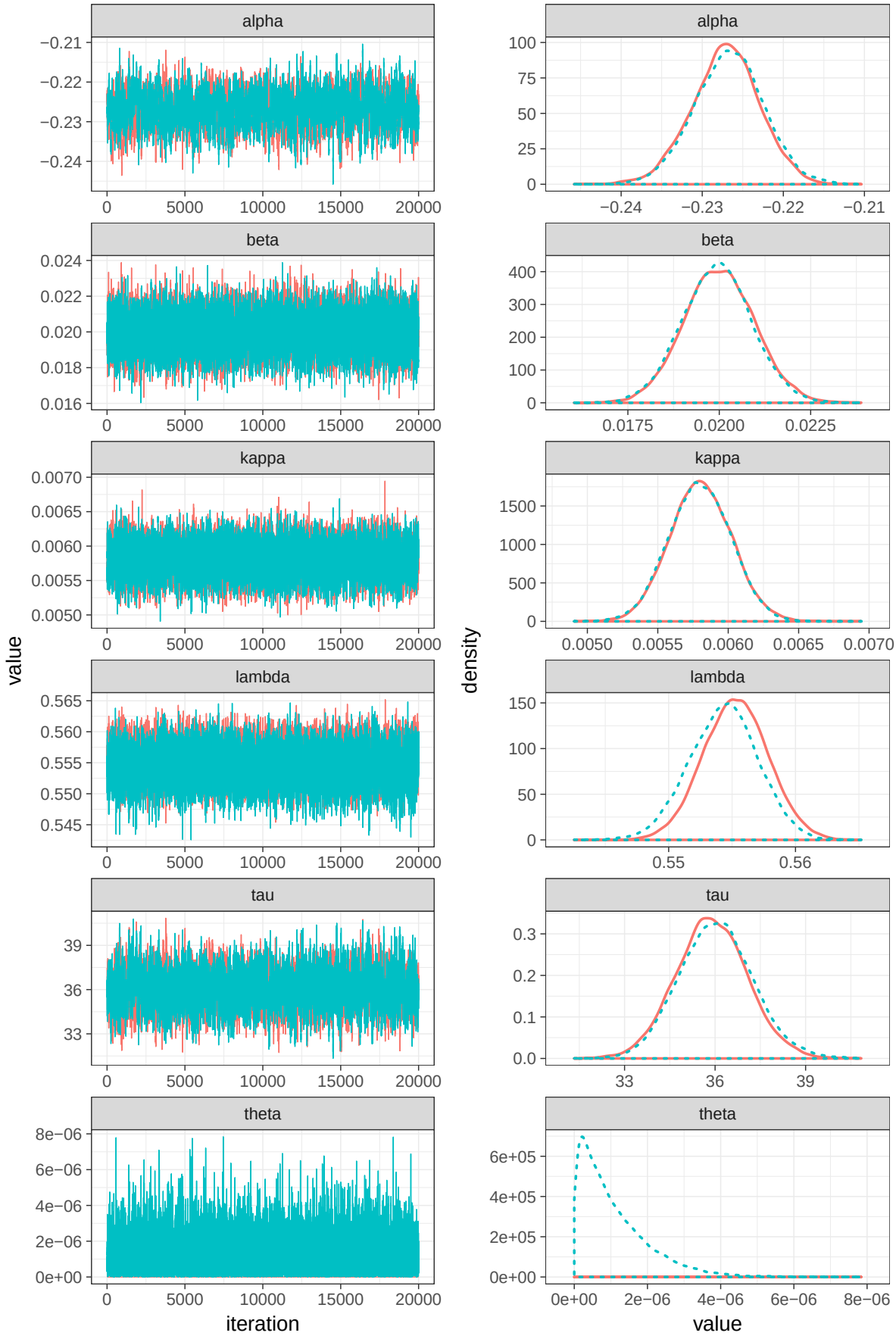

Figure 6: Traceplots (left) and posterior densities (right) of the parameters of models 0 (salmon, solid) and 1 (turquoise, dashed on the right) fitted to #gots7 data.

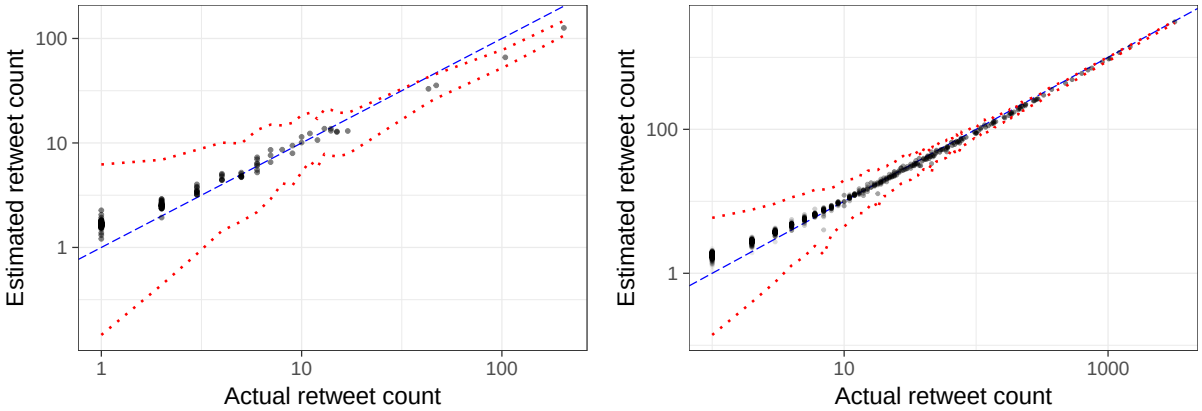

Figure 7: Estimated against actual retweet count for #thehandmaidstale (left) and #gots7 data (right) with 95% credible intervals (dotted, red) for the model selected by Gibbs variable selection. For multiple originals with the same actual retweet count, the widest credible interval is shown.

retweet count in Figure 7. On one hand, the model provides a reasonable fit to the #thehandmaidstale data, although small counts are overestimated while larger counts are overestimated. On the other hand, the model is doing a very good job for the larger #gots7 data. Interestingly, for either data set, the alternative model gives very close estimates of retweet count, which are therefore not plotted, to those predicted by the selected model.

7 Discussion

In this article we have proposed a Bayesian hierarchical model of hybrid processes, which is shown to model well the retweets of the originals of a specific hashtag. In the application to #thehandmaidstale data, the results of the model selection algorithm support the inclusion of θ for the exponential decay, whereas in the application to #gots7 data, the results suggest that it is sufficient to fit a special case of the proposed model, which is the hierarchical model of power law processes. Both sets of results are consistent with what Mathews et al. (2017) have found. Also, incorporating a censored regression model (with modification) in the proposed model allows us to explain retweet count by the follower count, which seems a natural candidate for driving retweet behaviour.

Whether the squared term $(x_i^*)^2$ should be included in δ_i can be examined by carrying out Gibbs variable selection (GVS) for κ, in the same way it is for θ in our inference and application. In fact, as GVS was originally developed for determining whether covariates in a linear regression model should be included or not, we could include terms of higher powers in the linear predictor, and perform GVS for the associated parameters (and κ and θ) simultaneously. However, as the fit of censored regression in Section 3 shows adequacy of using a quadratic polynomial, and as our focus is on the overall structure of the hierarchical

model instead of the particular form of the linear predictor, we confine the GVS to θ only in this paper.

While there are temporal fluctuations and bursty dynamics shown by, for example, Sanli and Lambiotte (2015), Mathiesen et al. (2013) Mollgaard and Mathiesen (2015), for their respective Twitter data sets, our results should not been seen as contradictory, for two reasons. First, their data were collected over several weeks, during which injections of interest due to external events were possible, while the data collection period was much shorter in both of our data sets, in which the generation of tweets and retweets was predominantly due to the broadcast of a TV series episode, the time of which was pre-specified. Second, our hierarchical model concerns the behaviour of information *diffusion*, in the form of retweets, which may be different from that of information *generation*, in the form of originals. Had the data been collected over a much longer time span, fitting the hybrid process to the originals, as we did in Section 2, might not be appropriate anymore. Therefore, even though the hybrid process may not be applicable to all kinds of online social media data of all timescales, it should still be useful to data regarding information diffusion within a time span of a single event with no apparent changepoints.

The parameters $\alpha, \beta, \kappa, \lambda, \tau$ and θ are assumed to be common to all hybrid processes of the originals. We argue that, for α, β, κ and τ, the variation in $e^{\delta_i} - 1$ is already captured by the covariate x_i^* together with ϵ_i, while for λ and θ, there is simply no extra information to assume otherwise. In the absence of other potentially useful covariates, it may not be useful to incorporate a hierarchical structure to any of these parameters, as doing so still attributes any unexplained variation to noise. The parsimony of the proposed model also makes it easy to simulate retweets given originals and follower counts. Given the parameters, we can directly simulate ϵ_i, δ_i and subsequently the retweet times using (5), (8) and (7), respectively. If appropriate, we can go one step further by first simulating the originals from a hybrid process with possibly different values of λ and θ, followed by the simulation of the retweets.

Another direction for further study is the generation of follower count. While both the follower count and the retweet count in our data sets empirically follow the power law, it is not used to characterise either of them. Instead of treating the former as a given covariate, it is possible to first draw the follower count from the power law, then draw retweet count given the follower count, again via the censored regression model.

Acknowledgement

Data supporting this publication is openly available under an 'Open Data Commons Open Database License'. Additional metadata are available at: http://dx.doi.org/10.17634/154300-57. Please contact

Newcastle Research Data Service at rdm@ncl.ac.uk for access instructions. This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) grant DERC: Digital Economy Research Centre (EP/M023001/1).

References

Antoniades, D. and Dovrolis, C. (2015), 'Co-evolutionary dynamics in social networks: a case study of Twitter', *Computational Social Networks* **2**(14).

Bao, P., Shen, H.-W., Jin, X. and Cheng, X.-Q. (2015), Modeling and predicting popularity dynamics of microblogs using self-excited Hawkes processes, *in* 'Proceedings of the 24th International Converence on World Wide Web', WWW '15 Companion, ACM, New York, NY, USA, pp. 9–10.

Bar-Lev, S. K., Lavit, I. and Reiser, B. (1992), 'Bayesian inference for the power law process', *Annals of the Institute of Statistical Mathematics* **44**(4), 623–639.

Barabási, A.-L. and Albert, R. (1999), 'Emergence of scaling in random networks', *Science* **286**(5439), 509–512.

Bhamidi, S., Steele, J. M. and Zaman, T. (2015), 'Twitter event networks and the Superstar model', *The Annals of Applied Probability* **25**(5), 2462–2502.

Carlin, B. P. and Chib, S. (1995), 'Bayesian model choice via Markov chain Monte Carlo methods', *Journal of Royal Statistical Society: Series B* **157**(3), 473–484.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002), 'On Bayesian model and variable selection using MCMC', *Statistics and Computing* **12**, 27–36.

Duane, J. T. (1964), 'Learning curve approach to reliability monitoring', *IEEE Transactions on Aerospace* **2**(2), 563–566.

Farajtabar, M., Wang, Y., Rodriguez, M. G., Li, S., Zha, H. and Song, L. (2015), Coevolve: a joint point process model for information diffusion and network co-evolution, *in* 'Advances in Neural Information Processing Systems 28', NIPS 2015.

Götz, M., Leskovec, J., McGlohon, M. and Faloutsos, C. (2009), Modeling blog dynamics, *in* 'ICWSM'.

Green, P. J. (1995), 'Reversible jump Markov chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82**(4), 711–732.

Hong, L., Doumith, A. S. and Davison, B. D. (2013), Co-factorization machines: modeling user interests and predicting individual decisions in Twitter, *in* 'Proceedings of the Sixth ACM International Conference on Web Search and Data Mining', WSDM '13, ACM, New York, NY, USA, pp. 557–566.
URL: *http://doi.acm.org/10.1145/2433396.2433467*

Kumar, S., Liu, H., Mehta, S. and Subramaniam, L. V. (2014), 'From tweets to events: exploring a scalable solution for Twitter streams', *ArXiv e-prints* .

Kumar, S., Liu, H., Mehta, S. and Subramaniam, L. V. (2015), Exploring a scalable solution to identifying events in noisy Twitter streams, *in* 'Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015', ASONAM '15, ACM, New York, NY, USA, pp. 496–499.

Li, H., Liu, J., Xu, K. and Wen, S. (2012), Understanding video propagation in online social networks, *in* 'Proceedings of the 2012 IEEE 20th International Workshop on Quality of Service'.

Li, P., Li, W., Wang, H. and Zhang, X. (2014), 'Modeling of information diffusion in Twitter-like social networks under information overload', *The Scientific World Journal* **2014**.

Lim, K. W., Chen, C. and Buntine, W. (2016), 'Twitter-network topic model: a full Bayesian treatment for social network and text modeling', *ArXiv e-prints* .

Mahmud, J., Chen, J. and Nichols, J. (2013), When will you answer this? Estimating response time in Twitter, *in* 'Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media', Association for the Advancement of Artificial Intelligence, pp. 697–700.

Masuda, N. and Rocha, L. E. C. (2017), 'A Gillespie algorithm for non-Markovian stochastic processes', *Society for Industrial and Applied Mathematics Review* **in press**.

Mathews, P., Mitchell, L., Nguyen, G. and Bean, N. (2017), The nature and origin of heavy tails in retweet activity, *in* 'Proceedings of the 26th International Conference on World Wide Web Companion', WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 1493–1498.
URL: *https://doi.org/10.1145/3041021.3053903*

Mathiesen, J., Angheluta, L., Ahlgren, P. T. H. and Jensen, M. H. (2013), 'Excitable human dynamics driven by extrinsic events in massive communities', *Proceedings of the National Academy of Sciences* **110**(43), 17259–17262.

Miotto, J. M., Kantz, H. and Altmann, E. G. (2017), 'Stochastic dynamics and the predictability of big hits in online videos', *Physical Review E* **95**, 032311.
URL: *https://link.aps.org/doi/10.1103/PhysRevE.95.032311*

Mollgaard, A. and Mathiesen, J. (2015), 'Emergent user behavior on Twitter modelled by a stochastic differential equation', *PLOS ONE* **10**(5), 1–12.
URL: *https://doi.org/10.1371/journal.pone.0123876*

Nishi, R., Takaguchi, T., Oka, K., Maehara, T., Toyoda, M., Kawarabayashi, K. and Masuda, N. (2016), 'Reply trees in Twitter: data analysis and branching process models', *Social Network Analysis and Mining* **6**(26), 1–13.

Perera, R. D. W., Anand, S., Subbalakshmi, K. P. and Chandramouli, R. (2010), Twitter analytics: Architecture, tools and analysis, *in* '2010 - MILCOM 2010 Military Communications Conference', pp. 2186–2191.

Pozdnoukhov, A. and Kaiser, C. (2011), Space-time dynamics of topics in streaming text, *in* 'Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks', LBSN '11, ACM, New York, NY, USA, pp. 1–8.

Sakaki, T., Okazaki, M. and Matsuo, Y. (2010), Earthquake shakes Twitter users: real-time event detection by social sensors, *in* 'Proceedings of the 19th International Conference on World Wide Web', WWW '10, ACM, New York, NY, USA, pp. 851–860.
URL: *http://doi.acm.org/10.1145/1772690.1772777*

Sanli, C. and Lambiotte, R. (2015), Local variation of collective attention in hashtag spike trains, *in* 'Modeling and Mining Temporal Interactions: Papers from the 2015 ICWSM Workshop', AAAI Press, Palo Alto, California.

Shen, H., Wang, D., Song, C. and Barabási, A.-L. (2014), Modeling and predicting popularity dynamics via reinforced Poisson processes, *in* 'Proceedings of the 28th AAAI Conference on Artificial Intelligence', Association for the Advancement of Artificial Intelligence, pp. 291–297.

Smid, H., Mast, P., Tromp, M., Winterboer, A. and Evers, V. (2011), Canary in a coal mine: monitoring air quality and detecting environmental incidents by harvesting Twitter, *in* 'CHI '11 Extended Abstracts on Human Factors in Computing Systems', CHI EA '11, ACM, New York, NY, USA, pp. 1855–1860.

Srijith, P. K., Lukasik, M., Bontcheva, K. and Cohn, T. (2017), 'Longitudinal modeling of social media with Hawkes process based on users and networks'.

Sutton, J., Spiro, E. S., Johnson, B., Fitzhugh, S., Gibson, B. and Butts, C. T. (2014), 'Warning tweets: serial transmission of messages during the warning phase of a disaster event', *Information, Communication & Society* **17**(6), 765–787.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica* **26**(1), 24–36.

Wu, M., Guo, J., Zhang, C. and Xie, J. (2011), Social media communication model research bases on Sina-Weibo, *in* 'Knowledge Engineering and Management: Proceedings of the Sixth International Converence on Intelligent Systems and Knowledge Enginnering, Shanghai, China, Dec 2011 (ISKE2011)', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 445–454.
URL: *https://doi.org/10.1007/978-3-642-25661-5_57*

Xie, J., Zhang, C. and Wu, M. (2011), Modeling microblogging communication based on human dynamics, *in* '2011 Eighth International Converence on Fuzzy Systems and Knowledge Discovery (FSKD)', Vol. 4, pp. 2290–2294.

Yakovlev, G., Rundle, J. B., Shcherbakov, R. and Turcotte, D. L. (2005), 'Inter-arrival time distribution for the non-homogeneous Poisson process', *ArXiv e-prints* .

Zhu, J., Xiong, F., Piao, D., Liu, Y. and Zhang, Y. (2011), Statistically modelling the effectiveness of disaster information in social media, *in* '2011 IEEE Global Humanitarian Technology Conference', pp. 431–436.

A Likelihood derivation

This appendix derives the likelihood of the hierarchical model for retweets in Section 4. We first consider the contribution of the i-th original regardless of the model choice, or the value of M equivalently. If $m_i = 0$, that is, no retweets have been observed, the contribution is simply

$$f(m_i, \boldsymbol{t}_i | x_i^*, \boldsymbol{\eta}_M, M) = \exp\left[-H_i(T)\right].$$

If $m_i > 0$, the contribution to the likelihood is

$$f(m_i, \boldsymbol{t}_i | x_i^*, \boldsymbol{\eta}_M, M) = \exp\left[-H_i(T)\right] \prod_{j=1}^{m_i} h_i(t_{ij}) \times \mathbf{1}_{\left\{\alpha+\beta x_i^* + \kappa\left(x_i^*\right)^2 + \epsilon_i > 0\right\}}.$$

The positivity of the linear predictor $\alpha + \beta x_i^* + \kappa \left(x_i^*\right)^2 + \epsilon_i$ is necessary for a positive intensity in the interval $[t_i, T]$, as retweets will occur with probability 0 if $h_i(t) = 0$ for $t_i \leq t \leq T$. The complete likelihood is

$$\begin{aligned} f(\boldsymbol{m}, \boldsymbol{t} | \boldsymbol{x}^*, \boldsymbol{\eta}_M, M) &= \prod_{i:m_i=0} f(m_i, \boldsymbol{t}_i | x_i^*, \boldsymbol{\eta}_M, M) \times \prod_{i:m_i>0} f(m_i, \boldsymbol{t}_i | x_i^*, \boldsymbol{\eta}_M, M) \\ &= \prod_{i:m_i=0} \exp\left[-H_i(T)\right] \times \prod_{i:m_i>0} \left\{ \exp\left[-H_i(T)\right] \prod_{j=1}^{m_i} h_i(t_{ij}) \times \mathbf{1}_{\{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)>0\}} \right\} \end{aligned}$$

$$= \prod_{i=1}^{n} \exp\left[-H_i(T)\right] \times \prod_{i:m_i>0} \mathbf{1}_{\{\delta_i(\alpha,\beta,\kappa,\epsilon_i,x_i^*)>0\}} \times \prod_{i:m_i>0} \prod_{j=1}^{m_i} h_i(t_{ij}). \tag{19}$$

From now on, we drop the arguments in $\delta_i(\alpha, \beta, \kappa, \epsilon_i, x_i^*)$ for convenience. When $M = 0$, that is, $\theta = 0$ and is removed from the parameter vector, substituting (1) and (7) into (19) yields

$$\begin{aligned}
&f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, \boldsymbol{\eta}_0, M = 0) \\
&= \prod_{i=1}^{n} \exp\left[-(e^{\delta_i} - 1)\frac{(T - t_i)^{1-\lambda}}{1-\lambda}\right] \times \prod_{i:m_i>0} \mathbf{1}_{\{\delta_i>0\}} \times \prod_{i:m_i>0} \prod_{j=1}^{m_i} (e^{\delta_i} - 1)(t_{ij} - t_i)^{-\lambda} \\
&= \exp\left[-\sum_{i=1}^{n} (e^{\delta_i} - 1)\frac{(T - t_i)^{1-\lambda}}{1-\lambda}\right] \times \prod_{i:m_i>0} \left[\mathbf{1}_{\{\delta_i>0\}} \left(e^{\delta_i} - 1\right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda}\right] \\
&= \exp\left(-(1-\lambda)^{-1} \sum_{i=1}^{n} \left[\left(e^{\delta_i} - 1\right)(T - t_i)^{1-\lambda}\right]\right) \times \prod_{i:m_i>0} \left[\mathbf{1}_{\{\delta_i>0\}} \left(e^{\delta_i} - 1\right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda}\right],
\end{aligned}$$

which is identical to (13). When $M = 1$, that is, $\theta > 0$, (19) becomes

$$\begin{aligned}
&f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, \boldsymbol{\eta}_1, M = 1) \\
&= \prod_{i=1}^{n} \exp\left[-(e^{\delta_i} - 1)\Gamma\left(1 - \lambda, \theta\left(T - t_i\right)\right)\theta^{\lambda-1}\right] \times \prod_{i:m_i>0} \mathbf{1}_{\{\delta_i>0\}} \\
&\quad \times \prod_{i:m_i>0} \prod_{j=1}^{m_i} (e^{\delta_i} - 1)(t_{ij} - t_i)^{-\lambda} e^{-\theta(t_{ij} - t_i)} \\
&= \exp\left[-\theta^{\lambda-1} \sum_{i=1}^{n} (e^{\delta_i} - 1)\Gamma\left(1 - \lambda, \theta(T - t_i)\right)\right] \times \prod_{i:m_i>0} \mathbf{1}_{\{\delta_i>0\}} \\
&\quad \times \prod_{i:m_i>0} \left[\left(e^{\delta_i} - 1\right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda}\right] \times \exp\left[-\theta \sum_{i:m_i>0} \sum_{j=1}^{m_i} (t_{ij} - t_i)\right] \\
&= \exp\left(-\theta^{\lambda-1} \sum_{i=1}^{n} \left[\left(e^{\delta_i} - 1\right)\Gamma\left(1 - \lambda, \theta(T - t_i)\right)\right]\right) \times \prod_{i:m_i>0} \mathbf{1}_{\{\delta_i>0\}} \\
&\quad \times \prod_{i:m_i>0} \left[\left(e^{\delta_i} - 1\right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda}\right] \times \exp\left[-\theta \sum_{i:m_i>0} \sum_{j=1}^{m_i} (t_{ij} - t_i)\right] \\
&= \exp\left(-\theta^{\lambda-1} \sum_{i=1}^{n} \left[\left(e^{\delta_i} - 1\right)\Gamma\left(1 - \lambda, \theta(T - t_i)\right)\right] - \theta \sum_{i:m_i>0} \sum_{j=1}^{m_i} (t_{ij} - t_i)\right) \\
&\quad \times \prod_{i:m_i>0} \left[\mathbf{1}_{\{\delta_i>0\}} \left(e^{\delta_i} - 1\right)^{m_i} \prod_{j=1}^{m_i} (t_{ij} - t_i)^{-\lambda}\right],
\end{aligned}$$

which is identical to (14).

B MWG algorithm

This appendix describes the MWG algorithm for $\boldsymbol{\eta}_M$ in step 2 of the algorithm outlined in Section 5. For any parameter or latent variable σ, scalar or vector, we define $\boldsymbol{\eta}_{M,-\sigma}$ to be the vector $\boldsymbol{\eta}_M$ without σ, and for notational convenience, we denote $\boldsymbol{\eta}_M$ with σ set to a value σ_0 by $(\sigma_0, \boldsymbol{\eta}_{M,-\sigma})$, regardless of the position of σ in $\boldsymbol{\eta}_M$.

Sampling α: Assume the current value is α. We propose α^* from a symmetrical proposal $q(\cdot|\alpha)$, and accept α^* with probability

$$\min\left(1, \frac{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\alpha^*, \boldsymbol{\eta}_{M,-\alpha}), M\right)\pi_\alpha(\alpha^*)}{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\alpha, \boldsymbol{\eta}_{M,-\alpha}), M\right)\pi_\alpha(\alpha)}\right),$$

where $\pi_\alpha(\cdot)$, along with the priors for other scalar parameters, is given by (15).

Sampling β: Assume the current value is β. We propose β^* from a symmetrical proposal $q(\cdot|\beta)$, and accept β^* with probability

$$\min\left(1, \frac{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\beta^*, \boldsymbol{\eta}_{M,-\beta}), M\right)\pi_\beta(\beta^*)}{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\beta, \boldsymbol{\eta}_{M,-\beta}), M\right)\pi_\beta(\beta)}\right).$$

Sampling κ: Assume the current value is κ. We propose κ^* from a symmetrical proposal $q(\cdot|\kappa)$, and accept κ^* with probability

$$\min\left(1, \frac{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\kappa^*, \boldsymbol{\eta}_{M,-\kappa}), M\right)\pi_\kappa(\kappa^*)}{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\kappa, \boldsymbol{\eta}_{M,-\kappa}), M\right)\pi_\kappa(\kappa)}\right).$$

Sampling λ: Assume the current value is λ. We propose λ^* from a symmetrical proposal $q(\cdot|\lambda)$, and accept λ^* with probability

$$\min\left(1, \frac{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\lambda^*, \boldsymbol{\eta}_{M,-\lambda}), M\right)\pi_\lambda(\lambda^*)\mathbf{1}_{\{\lambda^*<1\}}}{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\lambda, \boldsymbol{\eta}_{M,-\lambda}), M\right)\pi_\lambda(\lambda)\mathbf{1}_{\{\lambda<1\}}}\right).$$

Sampling τ: As we have assigned a conditional conjugate prior to τ, its full conditional posterior is given by

$$\tau|\boldsymbol{\eta}_{M,-\tau}, \boldsymbol{m}, \boldsymbol{t} \sim \text{Gamma}\left(a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2}\sum_{i=1}^{n}\epsilon_i^2\right).$$

Effectively, we can sample for τ via a Gibbs step conditional on just the latent variables $\boldsymbol{\epsilon}$.

Sampling ϵ_i $(i = 1, 2, \ldots, n)$ (for $M = 0$): Assume the current value is ϵ_i. We propose ϵ_i^* from a symmetrical proposal $q(.|\epsilon_i)$, and accept ϵ_i^* with probability

$$\min\left(1, \frac{g\left(\epsilon_i^*|\boldsymbol{\eta}_{0,-\boldsymbol{\epsilon}}\right)\pi_\epsilon(\epsilon_i^*|\tau)}{g\left(\epsilon_i|\boldsymbol{\eta}_{0,-\boldsymbol{\epsilon}}\right)\pi_\epsilon(\epsilon_i|\tau)}\right),$$

where
$$g\left(\epsilon_i|\boldsymbol{\eta}_{0,-\boldsymbol{\epsilon}}\right) = \begin{cases} \exp\left(-(e^{\delta_i}-1)\dfrac{(T-t_i)^{1-\lambda}}{1-\lambda}\right), & m_i = 0, \\ \exp\left(-(e^{\delta_i}-1)\dfrac{(T-t_i)^{1-\lambda}}{1-\lambda}\right)\left(e^{\delta_i}-1\right)^{m_i}\mathbf{1}_{\{\delta_i>0\}}, & m_i > 0, \end{cases}$$

and $\pi_\epsilon(\epsilon_i|\tau)$ is given by (5).

Sampling ϵ_i $(i = 1, 2, \ldots, n)$ (for $M = 1$): Using notation similar to those for $M = 0$, we accept the proposed ϵ^* with probability

$$\min\left(1, \frac{g\left(\epsilon_i^*|\boldsymbol{\eta}_{1,-\boldsymbol{\epsilon}}\right)\pi_\epsilon(\epsilon_i^*|\tau)}{g\left(\epsilon_i|\boldsymbol{\eta}_{1,-\boldsymbol{\epsilon}}\right)\pi_\epsilon(\epsilon_i|\tau)}\right),$$

where
$$g\left(\epsilon_i|\boldsymbol{\eta}_{1,-\boldsymbol{\epsilon}}\right) = \begin{cases} \exp\left(-(e^{\delta_i}-1)\Gamma\left(1-\lambda, \theta(T-t_i)\right)\theta^{\lambda-1}\right), & m_i = 0, \\ \exp\left(-(e^{\delta_i}-1)\Gamma\left(1-\lambda, \theta(T-t_i)\right)\theta^{\lambda-1}\right)\left(e^{\delta_i}-1\right)^{m_i}\mathbf{1}_{\{\delta_i>0\}}, & m_i > 0. \end{cases}$$

Sampling θ (for $M = 1$ only): Assume the current value is θ. We propose θ^* from a symmetrical proposal $q(\cdot|\theta)$, and accept θ^* with probability

$$\min\left(1, \frac{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\theta^*, \boldsymbol{\eta}_{1,-\theta}), M=1\right)\pi_\theta(\theta^*)\mathbf{1}_{\{\theta^*>0\}}}{f\left(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{x}^*, (\theta, \boldsymbol{\eta}_{1,-\theta}), M=1\right)\pi_\theta(\theta)\mathbf{1}_{\{\theta>0\}}}\right).$$

C RJMCMC algorithm

This appendix describes the algorithm of RJMCMC, which is an alternative to GVS for model selection outlined in Section 5. In additional to the notation defined in previous sections, we denote $p(m, m')$ as the jump probability from model m to model m'. This probability can be chosen to optimise the mixing of the algorithm, and has to be pre-specified for all pairs of m and m' (including $m = m'$). As the posterior model probabilities are theoretically not affected by the jump probabilities, in our application $p(0, 1)$ and $p(1, 0)$ are chosen to both be 0.5 for simplicity. Both $p(0, 0)$ and $p(1, 1)$ are subsequently 0.5 too, as $p(0, 0) + p(0, 1) = p(1, 0) + p(1, 1) = 1$. The algorithm is as follows:

1. The current values in the chain are $\boldsymbol{\eta}_M, \boldsymbol{\eta}_{\setminus M}$ and M.
2. Propose a jump to models M and $1-M$ with probabilities $p(M, M)$ and $p(M, 1-M)$, respectively.
3. If it is model M the jump is proposed to, update $\boldsymbol{\eta}_M$ using the MWG algorithm described in Appendix B, and the current value of M stays unchanged. If it is model $1 - M$ the jump is proposed to, go to the next two steps.
4. If $M = 0$ (and it is $M = 1$ the jump is proposed to), draw θ from its pseudoprior $\pi(\theta|M = 0)$, and write $\boldsymbol{\eta}_1' = (\boldsymbol{\eta}_0, \theta)$. If $M = 1$ (and it is $M = 0$ the jump is proposed to), write $\boldsymbol{\eta}_0' = \boldsymbol{\eta}_{1,-\theta}$, that is,

$\boldsymbol{\eta}_1$ with the value of θ dropped, so that $\boldsymbol{\eta}_1 = (\boldsymbol{\eta}_0^{'}, \theta)$.

5. Accept the proposed move to $\boldsymbol{\eta}_{1-M}^{'}$ and model $1-M$ with probability

$$\min\left(1, \frac{f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{\eta}_{1-M}^{'}, 1-M)\ \pi(\theta|1-M)\ \pi(1-M)\ p(1-M, M)}{f(\boldsymbol{m}, \boldsymbol{t}|\boldsymbol{\eta}_M, M)\ \pi(\theta|M)\ \pi(M)\ p(M, 1-M)}\right).$$

As in the GVS algorithm, both the pseudoprior $\pi(\theta|M=0)$ and the prior $\pi(\theta|M=1)$ of θ are involved in calculating the above probability, while the priors of the overlapping parameters are not required as they are the same under both models.