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

Clement Lee & Darren J Wilkinson

View supplementary material

Accepted author version posted online: 09 Mar 2019.

Submit your article to this journal

View Crossmark data

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, UK

[2]Open Lab, School of Computing, Newcastle University, UK

***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.*

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. As an introduction to the model for Twitter retweets proposed in this article, we will give a brief review for each of the two aspects.

1.1 Temporal dynamics

A common approach to modelling temporal dynamics is the use of one-dimensional non-homogeneous Poisson Processes (NHPP). 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, are 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 user 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 NHPP with this specific deterministic form of $h(t)$ is termed the *hybrid process*, which will be the cornerstone of our proposed model.

While relaxing the stationarity assumption of the HPP leads to the NHPP, the former, as one of the simplest point processes, has many other properties (and therefore has alternative characterisations), such as the interarrival times being independent and identically distributed (i.i.d.) exponential random ariables, the independence of the random variables of the numbers of points in disjoint intervals, and that points are i.i.d. over a bounded region of time or space conditional on the total number of points in that region. This allows defining broader classes of point processes by generalising the HPP in various ways. For example, relaxing the exponential distribution assumption for the interarrival times gives rise to the renewal processes. Further allowing that the interarrival times form a Markov chain leads to the Wold process. Alternative joint distributions being specified for the points in a bounded domain given the total number of points leads to finite point processes. All these processes can be found in, for example, Daley and Vere-Jones (2003).

Apart from generalising the HPP in various ways, extensions can be made for point processes in general. For example, if there are measurements (marks) associated with the locations of the points, the observed data can be modelled by a marked point process, in which the locations are modelled by a point process and the marks are further modelled by a distribution, possibly conditional on the locations. Also, while the (first-order) intensity function characterises the NHPP, higher-order intensity functions, which are defined in a similar way, can be used to complement the characterisation of other processes, such as the Markov point process, also known as the Gibbs process. For further references of these processes, please see, for example, Daley and Vere-Jones (2003) and Diggle (2013).

Another common approach to extending a NHPP is the incorporation of stochasticity in $h(t)$. This means that events are assumed to arise from a NHPP *conditional on h(t)*, which in turns arises from a separate stochastic

process. One prominent example is the log-Gaussian Cox process (Møller et al., 1998). Regarding the models for Twitter data relevant to our research, one example is the aforementioned model by Mollgaard and Mathiesen (2015). 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.

The SEHP is being used widely in different fields to capture triggering and clustering behaviour (Reinhart, 2018). One similar context to modelling Twitter data is earthquake modelling in seismology. Ogata (1988) developed the temporal epidemic-type aftershock sequence (ETAS) model, in which mainshocks arise from a NHPP called the background process, and aftershocks arise from a different NHPP triggered by the occurrence of a mainshock or an aftershock. While the Twitter original tweets and retweets can be seen as analogous to the mainshocks and aftershocks in the ETAS model, respectively, subtle differences to our proposed model exist and will be explained in Section 3. The spatio-temporal version of the ETAS model is reviewed in Chiodi and Adelfio (2017), with a focus on inference approaches.

One final aspect of extending a temporal NHPP or general point process is its spatial counterpart or spatio-temporal generalisation. The spatial component exists in some of the references above, such as Reinhart (2018) for the SEHP, but is not reviewed extensively here due to the temporal nature of our research. For further references, please see Cressie (1993), Illian et al. (2008), Daley and Vere-Jones (2008), Gelfand et al. (2010), Cressie and Wikle (2011), Diggle (2013) and Baddeley et al. (2015).

1.2 Topological Aspects

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.

1.3 Proposed model

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 retweets by a collection of NHPPs conditional on the existence of the original tweets, simply called the originals hereafter. The specific NHPP used in the modelling is based on the hybrid process as in 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 computationally 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. Introduced and explored in Section 2 are the two data sets, the subsets of which are fitted by the hybrid process and its special case. The hierarchical model for retweets is introduced in Section 3, with its likelihood derived. The inference algorithm and the model diagnostics procedure are outlined in Sections 4 and 5, respectively. The model is applied to the two previously introduced data sets in Section 6, as well as a simulated data set in Section 7 to show that the model is realistic and that the inference algorithm performs well. Section 8 concludes the article.

2 Data and exploratory analysis

A set of tweets (both originals and retweets) with the hashtag #thehandmaidstale was collected on 2017-06-14 for 21 hours after one episode of the relevant TV series was broadcasted. There are in total 2043 originals, 265 of which have been retweeted at least once during the observation period. The times of the originals are plotted on the left of Figure 1 using a histogram, with bins of 10 minutes. For these 265 originals there are 971 retweets, while the most retweeted original, called the top original hereafter, has been retweeted 204 times. For each original, the cumulative retweet count is plotted over time on the left of Figure 2, meaning that there are 265 trajectories of different colours in total.

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 7th season premiere of the TV series Game of Thrones was broadcasted. The numbers of originals, originals retweeted, total retweets and retweets of the top original are provided in Table 1 alongside the counterparts for #thehandmaidstale data reported above. The histogram of originals is plotted on the right of Figure 1, while the cumulative retweet counts over time for each original are plotted on the right of Figure 2.

2.1 Hybrid process and the Duane plot

Another way of visualising the temporality of the data is through the Duane plot (Duane, 1964). In order to do so we have to first introduce the hybrid process and the power law process. Consider a NHPP with intensity function

$$h(t) = \gamma t^{-\lambda} e^{-\theta t}, \qquad (1)$$

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

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

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:

$$f(t_1, t_2, \ldots, t_n \mid \lambda, \theta, \gamma) := \exp\left[-H(T)\right] \times \prod_{i=1}^{n} h(t_i) \qquad (2)$$

$$= \begin{cases} \exp\left[-\gamma \Gamma(1-\lambda, \theta T)\theta^{\lambda-1}\right] \times \prod_{i=1}^{n} \gamma t_i^{-\lambda} e^{-\theta t_i}, & \theta > 0, \\ \exp\left[-\gamma T^{1-\lambda} / (1-\lambda)\right] \times \prod_{i=1}^{n} \gamma t_i^{-\lambda}, & \theta = 0. \end{cases} \qquad (3)$$

The derivation of the likelihood for general temporal point processes, which also applies to (2), can be found in, for example, Daley and Vere-Jones (2003), Proposition 7.2.III. 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 some 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 $\mathrm{E}[N(t_i)]$, 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 $\mathrm{E}[N(t_i)] = H(t_i) = \dfrac{\gamma t_i^{1-\lambda}}{1-\lambda}$, and so, if it is equal to i, we have

$$
\frac{\gamma t_i^{1-\lambda}}{1-\lambda} = i \quad \Leftrightarrow \quad \frac{t_i}{i} = \frac{1-\lambda}{\gamma} t_i^{\lambda}
$$

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

(4)

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.

The Duane plot for the 2043 originals of #thehandmaidstale data is shown on the left of Figure 3, 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 (3) 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 -9422.9. 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 -9336.4. 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 3 is the Duane plot for the retweets of the top original (with 204 retweets) of #thehandmaidstale data, 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 (4) with $(\lambda, \gamma) = (\hat{\lambda}, \hat{\gamma})$, and its proximity provides further support to the adequacy of the power law process.

For the 25420 originals of #gots7 data, 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 -5545.8 and $(\hat{\lambda}, \hat{\gamma}) = (0.568, 25.104)$, which are used to obtain the theoretical line overlaid in the Duane plot in Figure 4. 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 to obtain maximised log-likelihood -5401.1 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 5, 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 generalise 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 4 through model selection.

The temporality of the originals and retweets aside, we are also interested in explaining the retweet count, which is the outcome of the process generating retweets, by the follower count, which is observed once the original is tweeted. We assume that the i-th original is tweeted at time s_i, when the author of which has $x_i^* \geq 0$ followers. At time T, the end of the observation period, there are $m_i \geq 0$ retweets observed for this original. Next, we define $m_i^* = \log(1 + m_i)$ and $x_i = \log(1 + x_i^*) - \frac{1}{n}\sum_{k=1}^{n}\log(1 + x_k^*)$ to be the "log" retweet count and "mean-centred" follower count, respectively. While x_i will be used as the covariate in the hierarchical modelling, the loose definition of the "log" retweet count is to ensure m_i^* is finite, which will only be used for exploratory purposes in this section. The scatterplots of m_i^* against x_i for the retweets of the aforementioned data sets are shown in Figure 6, indicating that linear regression may be appropriate on these scales. The full model in Section 3 adheres to such relationships as it essentially models the retweet count m_i according to a Poisson regression, in which the mean is proportional to the exponential of a linear predictor of the covariate x_i.

3 Hierarchical model and likelihood

For convenience, the terminology and notation in Section 2 are retained, but no model is assumed for how the *originals* are generated as it is not the concern of this section. Instead, modelled are the *retweets* of the i-th original ($i = 1, 2, \ldots, n$) observed in $[s_i, T]$. The times of these m_i retweets, denoted by $t_{i1}, t_{i2}, \ldots, t_{im_i}$ such that $s_i \leq t_{i1} \leq t_{i2} \leq \cdots \leq t_{im_i} \leq T$, are assumed to arise from a generalised hybrid process with intensity

$$h_i(t) = \phi e^{\delta_i}(t - s_i + \psi)^{-\lambda}e^{-\theta(t-s_i)}\boldsymbol{I}_{\{t \geq s_i\}}, \quad (5)$$

where $\lambda < 1, \phi > 0, \psi \geq 0, \theta \geq 0$, and $\boldsymbol{I}_{\{A\}}$ is the indicator function of event A. The cumulative intensity at time T follows directly:

$$H_i(T) = \int_0^T h_i(t)dt = \int_{s_i}^T \phi e^{\delta_i}(t - s_i + \psi)^{-\lambda} e^{-\theta(t-s_i)}dt$$

$$= \phi e^{\delta_i} \times \begin{cases} \left[(T - s_i + \psi)^{1-\lambda} - \psi^{1-\lambda}\right](1-\lambda)^{-1}, & \theta = 0, \quad (6) \\ \left[\Gamma(1-\lambda, \theta(T - s_i + \psi)) - \Gamma(1-\lambda, \theta\psi)\right]\theta^{\lambda-1}e^{\theta\psi}, & \theta > 0. \end{cases}$$

Several modifications from (1) can be observed in (5). First, the shift from t to $t - s_i$ in $h_i(t)$ is due to the process of retweets taking place relative to the time original i is tweeted. Second, while λ and θ are universal to the process of each original, δ_i is dependent on x_i and assumed to take the form

$$\delta_i = \beta x_i + \kappa x_i^2 + \epsilon_i, \quad (7)$$

where

$$\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \tau^{-1}), \qquad i = 1, 2, \ldots, n. \quad (8)$$

Third, there is no need for an intercept term in δ_i as it is embedded in ϕ, and ϕe^{δ_i} replaces γ in (1) as the scale component for the retweet count of the i-th original. Finally, the inclusion of ψ, along with other generalisations, is for comparison with the ETAS model, proposed by Ogata (1988) and mentioned in Section 1, later in this section.

Due to the nature of the NHPP, the retweet count m_i for the i-th original at time T, follows the Poisson distribution with mean $H_i(T) = \phi e^{\beta x_i + \kappa x_i^2 + \epsilon_i} \times$ (terms constant to x_i). Effectively and implicitly, a Poisson regression model is assumed for m_i by x_i, even though the former is not directly modelled. The use of the NHPP enables the retweet times to be modelled while simultaneously explaining the retweet count by the "mean-centred" follower count.

If we sum all intensities of the retweet process of individual originals, we obtain the overall intensity of a point process of all retweets:

$$h(t) := \sum_{i: t > s_i} h_i(t) = \sum_{i: t > s_i} \phi e^{\beta x_i + \kappa x_i^2 + \epsilon_i}(t - s_i + \psi)^{-\lambda} e^{-\theta(t-s_i)}. \quad (9)$$

This seems similar to the *conditional* intensity of the temporal ETAS model

$$h(t \mid \mathcal{H}_t) := \mu + \sum_{k:t>t_k} \phi e^{\beta(x_i - x_0)} \left(t - t_k + \psi\right)^{-\lambda}, \quad (10)$$

where $\mathcal{H}_t$ is the history of all events up to time t. The parameterisation is slightly different from that usually seen in the literature (Ogata, 1988; Chiodi and Adelfio, 2017; Reinhart, 2018), for the sake of easier alignment. The apparent differences include the presence of x_0 in (10), which however is usually given in modelling earthquake data, the inclusion of an extra parameter κ and a random effects term ϵ_i, which will be justified in our application in Section 6, and the additional exponential decay term over time $e^{-\theta(t-s_i)}$ suggested by Mathews et al. (2017). However, more important are the major differences in the underlying model structure. First, the ETAS model jointly models both the background events (originals in our case) and the triggered events (retweets) according to one point process, mainly because the nature of the events is not known prior to modelling, while in the proposed model concerned are the retweets *given the originals*. Such difference can be seen in (9) that μ is absent, which under the ETAS model means the originals are modelled by a HPP. Second, while the times s_i in (9) refer to the background events (originals) only, the times t_k in (10) refer to both type of events, hence the self-exciting nature of the process. This means under the ETAS model retweets can arise from other retweets, and different retweets essentially belong to a different and unobserved layer. On the other hand, under the proposed model, there is no self-excitation and only two layers of events exist, namely the originals from the background process, and the retweets that arise as offspring of the originals. Finally, under the ETAS model the layers, as well as the association between events of different layers, are unobserved and assumed by the model structure. Under the proposed model the relationships between the two completely known layers of events are observed and modelled accordingly.

Overall, our model deviates from the temporal ETAS model by removing the self-exciting nature of retweets. While it is possible to incorporate such structure, more information is required, such as the follower counts of the *retweeters*, in order to compute a less tractable likelihood. See, for example,

Reinhart (2018) for the complexity of the required computations. Rather, we utilise the most valuable information, which is the correspondance between the originals and the retweets, to compute a completely tractable likelihood, which will be shown towards the end of this section, under a model that elegantly encompasses a collection of NHPPs of retweets.

It is useful to define a few vectors for the derivations in the rest of this section. We write

$$
\begin{aligned}
&\boldsymbol{\eta}_0 := (\beta, \kappa, \lambda, \phi, \psi, \tau), \\
&\boldsymbol{\eta}_1 := (\beta, \kappa, \lambda, \phi, \psi, \tau, \theta), \\
&\boldsymbol{\epsilon} := (\epsilon_1, \epsilon_2, \ldots, \epsilon_n), \\
&\boldsymbol{m} := (m_1, m_2, \ldots, m_n), \\
&\boldsymbol{x} := (x_1, x_2, \ldots, x_n), \\
&\boldsymbol{s} := (s_1, s_2, \ldots, s_n), \\
&\boldsymbol{t}_i := (t_{i1}, t_{i2}, \ldots, t_{im_i}), \quad i = 1, 2, \ldots, n,
\end{aligned}
$$

and

$$
\boldsymbol{t} := \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}.
$$

The two vectors $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ correspond to the generalised power law and hybrid processes, respectively. When there is no confusion, we simply write $\boldsymbol{\eta}$ to represent this vector of scalar parameters. While the four vectors $\boldsymbol{\epsilon}$, $\boldsymbol{m}$, $\boldsymbol{x}$ and $\boldsymbol{s}$ are all of length n, $\boldsymbol{\epsilon}$ is a vector of latent variables whereas the others are given as data/covariates. Finally, $\boldsymbol{t}_i (i = 1, 2, \ldots, n)$ is the vector of retweet times of the i-th original, while $\boldsymbol{t}$ is the collection of retweet times of *all* originals. The length of $\boldsymbol{t}$, or equivalently the total number of retweets, is denoted by $m = \sum_{i=1}^{n} m_i$.

Before writing out the likelihood, we introduce a parameter *M*, which can take value 0 or 1, to represent model *choice*. When *M* = 0, the hierarchical model of the generalised 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 generalised 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 4. Our algorithm requires the likelihood for each model as a function of $\boldsymbol{\eta}_M$:

$$
\begin{aligned}
&f(\boldsymbol{m}, \boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\epsilon}, \boldsymbol{\eta}_0, M = 0) \\
&= \exp\left(-\frac{\phi}{1-\lambda} \sum_{i=1}^{n} e^{\beta x_i + \kappa x_i^2 + \epsilon_i} \left[(T - s_i + \psi)^{1-\lambda} - \psi^{1-\lambda}\right]\right) \\
&\quad \times \phi^m \exp\left(\sum_{i=1}^{n} m_i \left[\beta x_i + \kappa x_i^2 + \epsilon_i\right]\right) \times \prod_{i:m_i>0} \prod_{j=1}^{m_i} \left(t_{ij} - s_i + \psi\right)^{-\lambda},
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
&f(\boldsymbol{m}, \boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\epsilon}, \boldsymbol{\eta}_1, M = 1) \\
&= \exp\left(-\phi \theta^{\lambda-1} e^{\theta \psi} \sum_{i=1}^{n} e^{\beta x_i + \kappa x_i^2 + \epsilon_i} \left[\Gamma\left(1-\lambda, \theta(T - s_i + \psi)\right) - \Gamma\left(1-\lambda, \theta \psi\right)\right]\right) \\
&\quad \times \phi^m \exp\left(\sum_{i=1}^{n} m_i \left[\beta x_i + \kappa x_i^2 + \epsilon_i\right]\right) \\
&\quad \times \prod_{i:m_i>0} \prod_{j=1}^{m_i} \left(t_{ij} - s_i + \psi\right)^{-\lambda} \times \exp\left(-\theta \sum_{i:m_i>0} \sum_{j=1}^{m_i} \left(t_{ij} - s_i\right)\right).
\end{aligned}
$$
(12)

The derivations of (11) and (12) are detailed in Appendix A. Note that, even though τ is seen in neither (11) nor (12) because of independence between τ and the data conditional on ϵ, it is included in $\boldsymbol{\eta}_M$ for notational convenience.

4 Inference and the Bayes factor

The presence of the latent variables ϵ 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:

$$\beta \quad \sim \quad \mathrm{N}\left(\mu_\beta = 0, \tau_\beta^{-1} = 10^4\right),$$

$$\kappa \quad \sim \quad \mathrm{N}\left(\mu_\kappa = 0, \tau_\kappa^{-1} = 10^4\right),$$

$$(1-\lambda) \quad \sim \quad \mathrm{Gamma}\left(a_\lambda = 1, b_\lambda = 0.001\right),$$

$$\phi \quad \sim \quad \mathrm{Gamma}\left(a_\phi = 1, b_\phi = 0.001\right), \qquad (13)$$

$$\psi \quad \sim \quad \mathrm{Gamma}\left(a_\psi = 1, b_\psi = 0.001\right),$$

$$\tau \quad \sim \quad \mathrm{Gamma}\left(a_\tau = 1, b_\tau = 0.001\right),$$

$$\theta \quad \sim \quad \mathrm{Gamma}\left(a_\theta = 1, b_\theta = 0.001\right), \quad (\text{only for } M = 1)$$

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 $Y \sim \mathrm{Gamma}\left(a_Y, b_Y\right)$. As the parameter space is not the same for η_0 and η_1, we denote $\eta_{\backslash M}$ as the subset of η_M not in η_{1-M}, which means $\eta_{\backslash 0} = \theta$ and $\eta_{\backslash 1} = \{\}$, the null set. Assuming conditional independence of η_M and $\eta_{\backslash M}$ given M, the joint posterior of $\epsilon, \eta_M, \eta_{\backslash M}$ and M is

$$\pi(\epsilon, \eta_M, \eta_{\backslash M}, M \mid m, t, x, s) \propto \pi(m, t, \epsilon, \eta_M, \eta_{\backslash M}, M \mid x, s)$$
$$= f(m, t \mid x, s, \epsilon, \eta_M, M) \times \pi(\epsilon \mid \eta_M) \times \pi(\eta_M \mid M) \times \pi(\eta_{\backslash M} \mid M) \times \pi(M),$$
$$(14)$$

where $\pi(\epsilon \mid \eta_M)$ can be simplified to $\pi(\epsilon \mid \tau)$ and is given by (8), while the last component $\pi(M)$ is the prior model probability. The "true" prior of η_M under M is given by

$$\pi(\eta_M \mid M) = \begin{cases} \pi_\beta(\beta)\pi_\kappa(\kappa)\pi_\lambda(\lambda)\pi_\psi(\psi)\pi_\phi(\phi)\pi_\tau(\tau), & M = 0, \\ \pi_\beta(\beta)\pi_\kappa(\kappa)\pi_\lambda(\lambda)\pi_\psi(\psi)\pi_\phi(\phi)\pi_\tau(\tau)\pi_\theta(\theta \mid M = 1), & M = 1, \end{cases}$$

where the components are given by (13). For $M = 1$, the pseudoprior $\pi(\eta_{\backslash M} \mid M)$ vanishes as $\eta_{\backslash 1} = \{\}$, while for $M = 0$, the pseudoprior can be written as $\pi_\theta(\theta \mid M = 0)$ equivalently. We proceed to draw samples of $(\epsilon, \eta_M, \eta_{\backslash M}, M)$ 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 $\epsilon, \eta_M, \eta_{\backslash M}$ and M.

2. Draw $\boldsymbol{\eta}_M$ from its conditional posterior, with density proportional to

$f(\boldsymbol{m},\boldsymbol{t}\,|\,\boldsymbol{x},s,\epsilon,\boldsymbol{\eta}_M,M)\times\pi(\epsilon\,|\,\tau)\times\pi(\boldsymbol{\eta}_M\,|\,M)$, 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. Draw ϵ from its conditional posterior, with density proportional to

$f(\boldsymbol{m},\boldsymbol{t}\,|\,\boldsymbol{x},s,\epsilon,\boldsymbol{\eta}_M,M)\times\pi(\epsilon\,|\,\tau)$, by the same MWG alogithm in Appendix B. Denote the value by $\epsilon^{'}$.

4. If $M = 0$, draw θ from its pseudoprior $\pi_\theta(\theta\,|\,M=0)$. 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.

5. Draw M from $\pi(M\,|\,\boldsymbol{m},\boldsymbol{t},\boldsymbol{x},s,\epsilon,\boldsymbol{\eta}_M,\boldsymbol{\eta}_{\backslash M})$, its conditional posterior distribution. 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 (14),

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

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

The pseudoprior $\pi_\theta(\theta\,|\,M=0)$ is chosen to be close to the marginal posterior of θ under the competing model, denoted by $\pi_\theta(\theta\,|\,M=1,\boldsymbol{m},\boldsymbol{t},\boldsymbol{x},s)$, for the sake of optimisation (Dellaportas et al., 2002; Carlin and Chib, 1995). It 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 $\pi_\theta(\theta\,|\,M=0)$ and the prior $\pi_\theta(\theta\,|\,M=1)$ are involved in step 5.

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 $\hat{\pi}(M \mid \boldsymbol{m}, \boldsymbol{t})$, 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:

$$B_{10} = \frac{\hat{\pi}(M=1 \mid \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{s})}{\hat{\pi}(M=0 \mid \boldsymbol{m}, \boldsymbol{t}, \boldsymbol{x}, \boldsymbol{s})} / \frac{\pi(M=1)}{\pi(M=0)}. \qquad (15)$$

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.

5 Model diagnostics

In this section, we augment the function arguments of the cumulative intensity in (6) by writing $H_i(T; x_i, s_i, \epsilon_i, \boldsymbol{\eta}) := H_i(T)$, where x_i, s_i, ϵ_i and $\boldsymbol{\eta}$ are included whenever necessary. Under the proposed model, the random variable of the retweet count m_i at time *T* is Poisson distributed with mean (and variance) $H_i(T; x_i, s_i, \epsilon_i, \boldsymbol{\eta})$ which, according to (6), is

$$\mathrm{E}_{m_i}\left[m_i \mid x_i, s_i, \epsilon_i, \boldsymbol{\eta}, T \right] := H_i(T; x_i, s_i, \epsilon_i, \boldsymbol{\eta}) = H_i(T; x_i, s_i, 0, \boldsymbol{\eta}) \times e^{\epsilon_i}. \qquad (16)$$

The second equality can be seen by substituting (7) into (6). As ϵ_i is $\mathrm{N}(0, \tau^{-1})$ distributed *apriori*, e^{ϵ_i} is log-normally distributed with mean $e^{0.5/\tau}$ and variance $e^{2/\tau} - e^{1/\tau}$. This enables us to obtain the expectation and variance of the expected retweet count in (16) *with respect to* ϵ_i:

$$\mathrm{E}_{\epsilon_i}\left[\mathrm{E}_{m_i}\left[m_i \big| x_i, s_i, \epsilon_i, \boldsymbol{\eta}, T \right] \right] = \mathrm{E}_{\epsilon_i}\left[H(T; x_i, s_i, 0, \boldsymbol{\eta}) \times e^{\epsilon_i} \right]$$
$$= H_i(T; x_i, s_i, 0, \boldsymbol{\eta}) \times e^{0.5/\tau},$$
$$\mathrm{Var}_{\epsilon_i}\left[\mathrm{E}_{m_i}\left[m_i \big| x_i, s_i, \epsilon_i, \boldsymbol{\eta}, T \right] \right] = \mathrm{Var}_{\epsilon_i}\left[H(T; x_i, s_i, 0, \boldsymbol{\eta}) \times e^{\epsilon_i} \right]$$
$$= \left(H_i(T; x_i, s_i, 0, \boldsymbol{\eta}) \right)^2 \times \left(e^{2/\tau} - e^{1/\tau} \right).$$

These two quantities can be used to obtain the χ^2 discrepancy, which is a goodness-of-fit measure advocated by Gelman, Meng and Stern (1996):

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(m_i - \mathrm{E}_{\epsilon_i}\left[\mathrm{E}_{m_i}\left[m_i | x_i, s_i, \epsilon_i, \boldsymbol{\eta}, T\right]\right]\right)^2}{\mathrm{Var}_{\epsilon_i}\left[\mathrm{E}_{m_i}\left[m_i | x_i, s_i, \epsilon_i, \boldsymbol{\eta}, T\right]\right]}$$

(17)

$$= \left(e^{2/\tau} - e^{1/\tau}\right)^{-1} \sum_{i=1}^{n} \left(\frac{m_i}{H_i\left(T; x_i, s_i, 0, \boldsymbol{\eta}\right)} - e^{0.5/\tau}\right)^2.$$

This χ^2 discrepancy is directly computable using the samples of $\boldsymbol{\eta}$ and ϵ from the MCMC outlined in Section 4. At each iteration, using the observed retweet counts as $\boldsymbol{m} = (m_1, m_2, \ldots, m_n)$ and the current values of $\boldsymbol{\eta}$, the actual discrepancy, denoted by χ^2_{act}, can be obtained. On the other hand, using the current values of $\boldsymbol{\eta}$ and ϵ_i, we can simulate the retweet count, for the i-th original, from the Poisson distribution with mean $\mathrm{E}_{m_i}\left[m_i | x_i, s_i, \epsilon_i, \boldsymbol{\eta}, T\right] = H_i(T; x_i, s_i, \epsilon_i, \boldsymbol{\eta})$. Plugging the whole set of simulated retweet counts as $\boldsymbol{m}$ and the current values of $\boldsymbol{\eta}$ into (17) yields the simulated discrepancy, denoted by χ^2_{sim}. Finally, comparing the two sets of discrepancies will help us determine if there are any inadequecies of the model fit. This can be achieved by plotting χ^2_{sim} against χ^2_{act} and computing the associated posterior predictive p-value, which is the empirical proportion of $\chi^2_{\mathrm{sim}} > \chi^2_{\mathrm{act}}$. Both of these will be presented in Section 6. For details of diagnostics for Bayesian hierarchical models in general, please see, for example, Gelman et al. (1996) and Section 2.2.2 of Cressie and Wikle (2011).

6 Application

Both the model-specific algorithms in Appendix B and the model selection algorithm in 4 are applied to the two data sets with different hashtags. All MCMC runs were performed on a Linux machine with Intel Core i5-4690S Processor (3.2GHz). 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 7, and the computation times are reported in Table 2. 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 artifically to be 10^{-9} and $1 - 10^{-9}$, respectively. They are chosen this way not to represent our prior belief in the models, but to ensure sufficient mixing between the two states of M in the chain. Model 0 is selected for 7584 times out of 20000 iterations, meaning that the Bayes factor B_{10} in (15) is estimated to be

$\frac{12416}{7584} / \frac{1.0 - 10^{-9}}{10^{-9}} = 1.637 \times 10^{-9}$. The RJMCMC algorithm gives a similar

estimate of $B_{10} = \frac{12402}{7598} / \frac{1.0 - 10^{-9}}{10^{-9}} = 1.632 \times 10^{-9}$. So, for the #thehandmaidstale

data set, which consists of tweets for over 21 hours, the generalised power law process hierarchical model is more appropriate. Such findings are different from 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. For each of the two model-specific algorithms and the model selection algorithm, a chain of 20000 iterations is obtained, upon thinning of 1000, after discarding the first 1000000 as burn-in. The traceplots and the posterior densities for the parameters are plotted in Figure 8. The proximity of the posterior densities for all parameters other than θ is similar to that for the #thehandmaidstale data, suggesting that the inclusion of θ does not improve fit much. 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^{-10} and $1 - 10^{-10}$, respectively, model 0 is selected for 7344 times out of 20000 iterations in total, meaning that the Bayes factor is

estimated to be $B_{10} = \frac{12656}{7344} / \left(\frac{1 - 10^{-10}}{10^{-10}} \right) = 1.723 \times 10^{-10}$. The RJMCMC algorithm

gives a similar estimate of $B_{10} = \frac{12727}{7273} / \left(\frac{1 - 10^{-10}}{10^{-10}} \right) = 1.75 \times 10^{-10}$, meaning that

model 0 is highly favoured. That the genearlised 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.

As outlined in Section 5, at each iteration of the MCMC, a value can be drawn from the Poisson distribution with mean $H_i(T; x_i, s_i, \epsilon_i, \boldsymbol{\eta})$ as the simulated (or predicted) retweet count of the i-th original $(i = 1, 2, \ldots, n)$. Across all iterations, the simulated values can be seen as a sample from the *posterior predictive distribution* of the random variable of the retweet count. As a preliminary way of examining the goodness-of-fit of our model, the posterior predictive mean of the retweet count, under the selected model 0 for both data sets, is plotted against the actual retweet count in Figure 9 with 95% predictive intervals. That some points seemingly form a vertical line is because there are multiple originals with the same actual retweet count. While small counts are slightly under-predicted, our model is doing a very good job for moderate to large counts for both data sets. For either data set, the alternative model 1 gives very close predictions of the retweet count, which are therefore not plotted, to those predicted by the selected model 0.

Utilising the simulated retweet counts, summaries of which are presented in Figure 9, the formal diagnostic procedures outlined in Section 5 are carried out for both data sets. The simulated discrepancies χ^2_{sim} are plotted against the actual discrepancies χ^2_{act} in Figure 10. The hovering of the points around the line $y = x$ shows no apparent lack-of-fit issues. This is supported by calculating the posterior predictive p-values, which are 0.424 and 0.436, for #thehandmaidstale and #gots7 data, respectively.

7 Simulation and backfitting

For each of the two data sets, we simulate the retweet times from the NHPP with intensity (5), based on the original times (s) and the "mean-centred" follower counts (x). The posterior means of the parameters in the respective selected models (model 0 for both data sets) are used as their true values in the simulation. The cumulative retweet counts over time for the simulated data are plotted in Figure 11. Comparing with their counterparts in Figure 2, the simulated data show a high degree of similarity in the overall pattern of retweet growth. Furthermore, the frequencies of retweet counts are equally highly similar between the simulated data and the actual data, as shown in

Figure 12. Both visualisations indicate that our proposed model is very capable of generating realistic retweet times that look like the observed data.

Not only is the model able to produce realistic realisations, but its inference procedure is also working properly and can recover the true parameter values. Specifically, we applied the MWG algorithm for the individual models (Appendix B) to the simulated data shown on the left of Figure 11. For each parameter (other than θ), its true value lies within the 95% credible interval of the corresponding marginal posterior distribution, as shown in Figure 13. Furthermore, the subsequent model selection overwhelming chooses model 0, which is the true model, over model 1. The Bayes factors B_{10} according to GVS and RJMCMC are 1.829×10^{-8} and 1.834×10^{-8}, respectively.

8 Discussion

In this article we have proposed a Bayesian hierarchical model of generalised hybrid processes, which is shown to model well the retweets of the originals of a specific hashtag. The application to both #thehahndmaidstale and #gots7 data suggests that it is sufficient to fit a special case of the proposed model, which is the hierarchical model of generalised power law processes. Also, incorporating an original-specific scale δ_i 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 GVS or RJMCMC 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 overall fit is adequate while 0 is excluded from the support of the marginal posterior of κ for both data sets, 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 article.

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 generalised hybrid process to the originals, as we did in Section 2, might not be appropriate anymore. Therefore, even though the generalised 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 $\beta, \kappa, \lambda, \phi, \psi, \tau$ and θ are assumed to be common to all generalised hybrid processes of the originals. We argue that, for β, κ, ϕ and τ, the between-original variation is already captured by the covariate x_i together with ϵ_i, while for ψ there is simply no extra information to assume otherwise. It might however be useful to incorporate a hierarchical structure to λ and θ simultaneously, to allow variation in the rate of decay of information diffusion between originals of the same hashtag. 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 (8), (7) and (5), respectively. If appropriate, we can go one step further by first simulating the originals from a more general point process, 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, via the implicit Poisson regression model.

Acknowledgements

References

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

Baddeley, A., Rubak, E. and Turner, R. (2015), *Spatial Point Patterns: Methodology and Applications with R*, Interdisciplinary Statistics Series, Chapman & Hall/CRC.

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.

Chiodi, M. and Adelfio, G. (2017), 'Mixed non-parametric and parametric estimation techniques in R package etasFLP for earthquakes' description', *Journal of Statistical Software* **76**, 1–29.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Wiley Series in Probability and Statistics, John Wiley & Sons.

Daley, D. J. and Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, Probability and Its Applications, Springer.

Daley, D. J. and Vere-Jones, D. (2008), *An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure*, Probability and Its Applications, Springer.

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

Diggle, P. J. (2013), *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, Monographs on Statistics and Applied Probability, Chapman & Hall.

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.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P., eds (2010), *Handbook of Spatial Statistics*, Handbooks of Modern Statistical Methods, Chapman & Hall/CRC.

Gelman, A., Meng, X. and Stern, H. (1996), 'Posterior predictive assessment of model fitness via realized discrepancies', *Statistica Sinica* **6**, 733–807.

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.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008), *Statistical Analysis and Modelling of Spatial Point Patterns*, John Wiley & Sons, Ltd.

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.

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.

Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998), 'Log Gaussian Cox processes', *Scandinavian Journal of Statistics* 25(3), 451–482.

Mollgaard, A. and Mathiesen, J. (2015), 'Emergent user behavior on Twitter modelled by a stochastic differential equation', *PLOS ONE* **10**(5), 1–12.

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.

Ogata, Y. (1988), 'Statistical models for earthquake occurrences and residual analysis for point processes', *Journal of the American Statistical Association* **83**(401), 9–27.

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.

Reinhart, A. (2018), 'A review of self-exciting spatio-temporal point processes and their applications', *Statistical Science* **33**(3), 299–318.

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.

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, *in* 'Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining', ACM.

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.

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.

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.
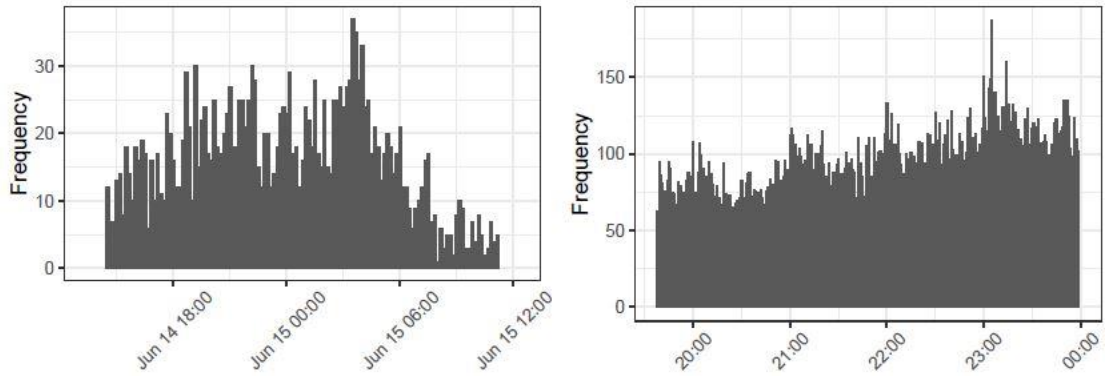
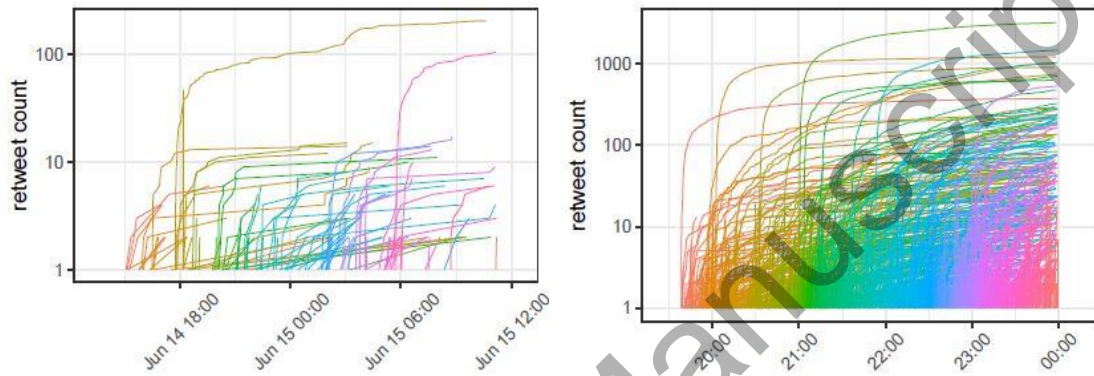**Fig. 1** Histogram of originals over time for the #thehandmaidstale (left) and #gots7 (right) data.

Fig. 2 Cumulative retweet counts over time for #thehandmaidstale (left) and #gots7 (right) data. Each trajectory is of a different colour and represents the growth of retweets of one individual original.
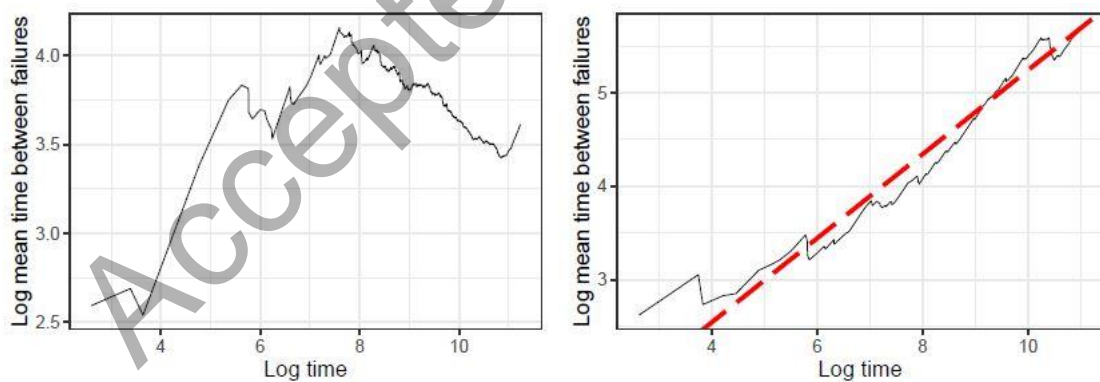

Fig. 3 Duane plots of originals (left), where time is relative to the start of observation period, and retweets of the top original (right), where time is relative to when the original is tweeted, for #thehandmaidstale data. Overlaid is the theoretical line (dashed) according to (4) with $(\lambda, \gamma) = (\hat{\lambda}, \hat{\gamma})$.
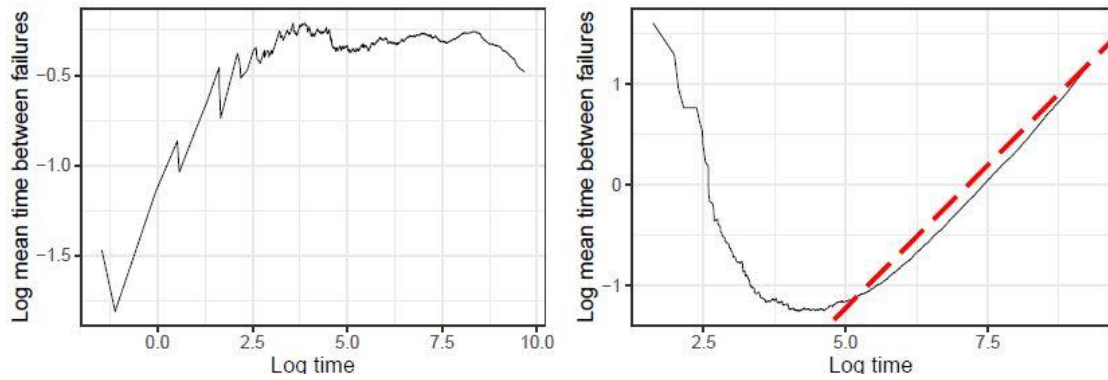
Fig. 4 Duane plots of originals (left) and retweets of the top original (right) for the #gots7 data, overlaid by the theoretical line (dashed) according to (4).
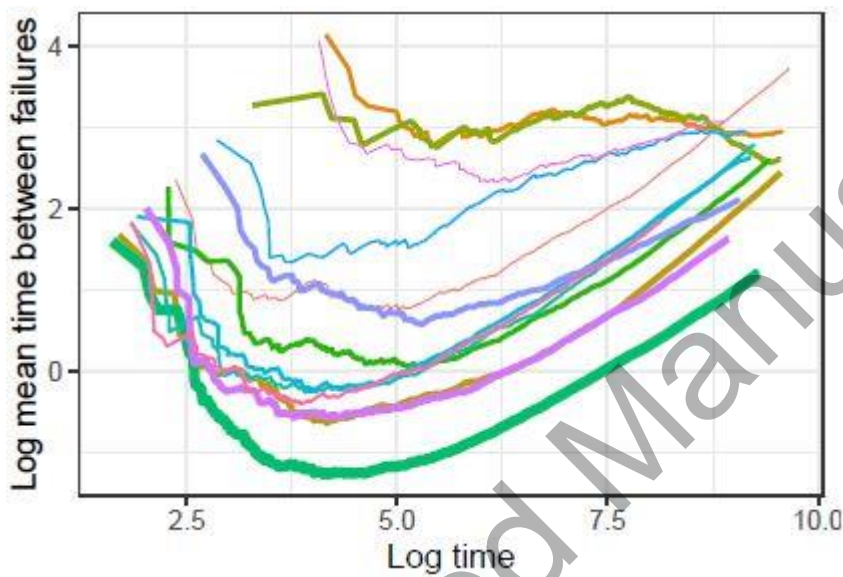
Fig. 5 Duane plots of retweets of the top 13 originals for the #gots7 data. The thicker the line is, the more retweets the original has.
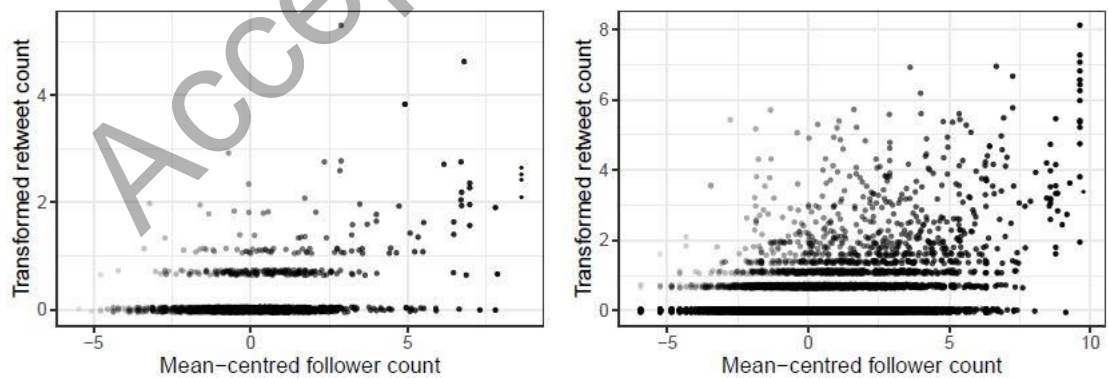
Fig. 6 Log retweet count against mean-centred follower count for #thehandmaidstale data (left) and #gots7 data (right).
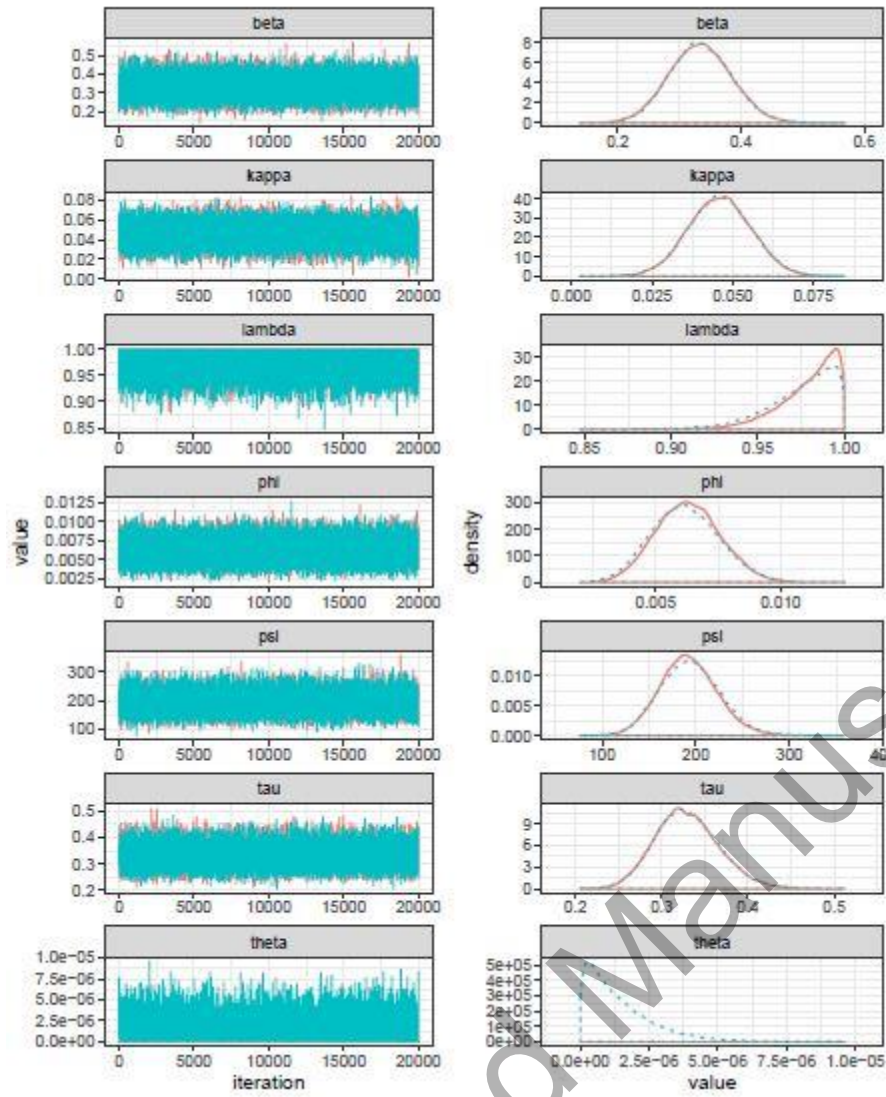
Fig. 7 Traceplots (left) and posterior densities (right) of the parameters of models 0 (salmon, solid) and 1 (turquoise, dashed on the right) fitted to #thehandmaidstale data.
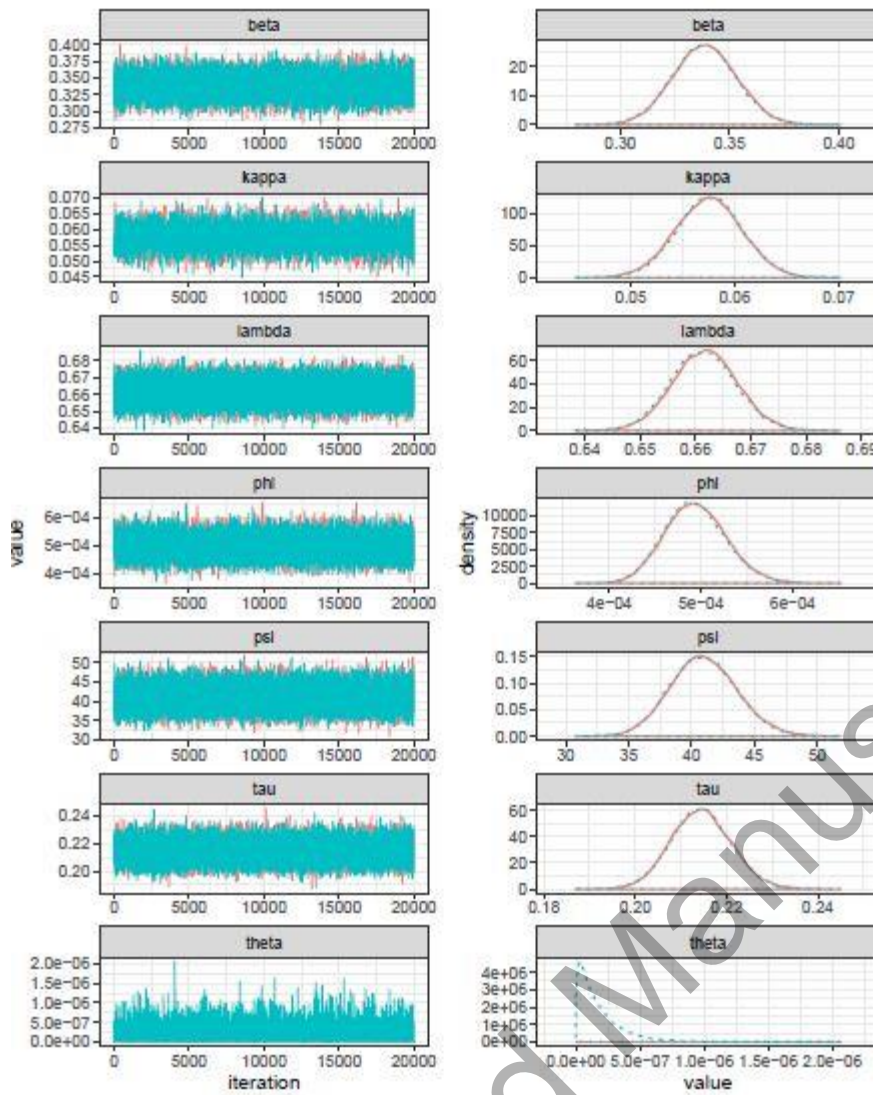
Fig. 8 Traceplots (left) and posterior densities (right) of the parameters of models 0 (salmon, solid) and 1 (turquoise, dashed on the right) fitted to #gots7 data.
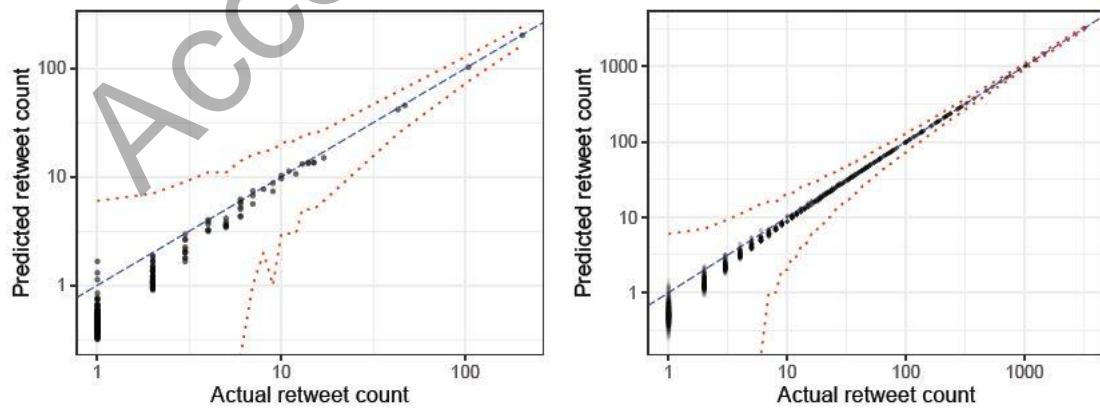
Fig. 9 Predicted against actual retweet count for #thehandmaidstale (left) and #gots7 data (right) with 95% prediction intervals (dotted, red) for the model selected by GVS. For multiple originals with the same actual retweet count, the widest prediction interval is shown. The blue dashed line is the line $y = x$.
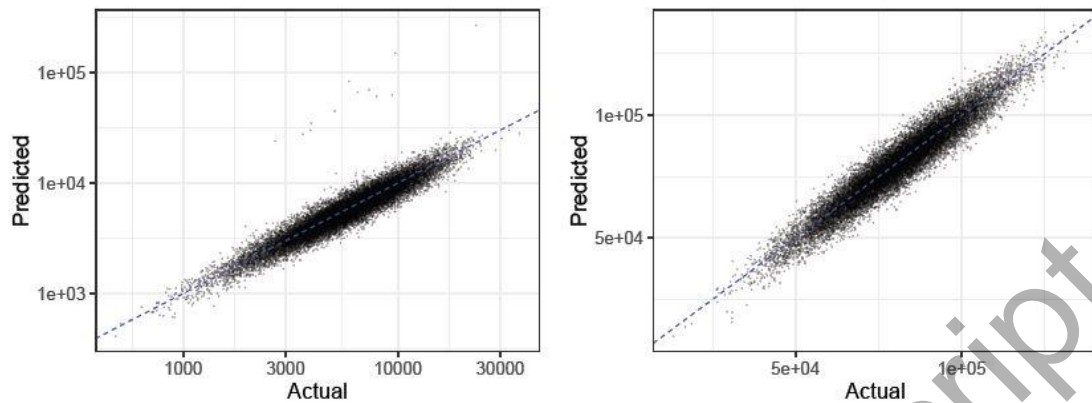

Fig. 10 Simulated against actual χ^2 discrepancies for #thehandmaidstale (left) and #gots7 (right) data. The blue dashed line is the line $y = x$.
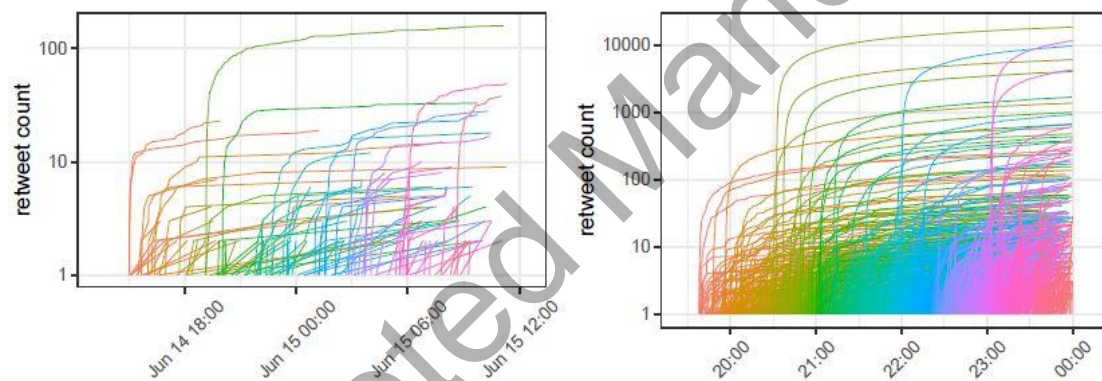

Fig. 11 Cumulative retweet counts over time for simulated data corresponding to #thehandmaidstale (left) and #gots7 (right) data. The posterior means of the parameters under model 0 are used as the true values in the simulation. Each trajectory is of a different colour and represents the growth of (simulated) retweets of one individual original.
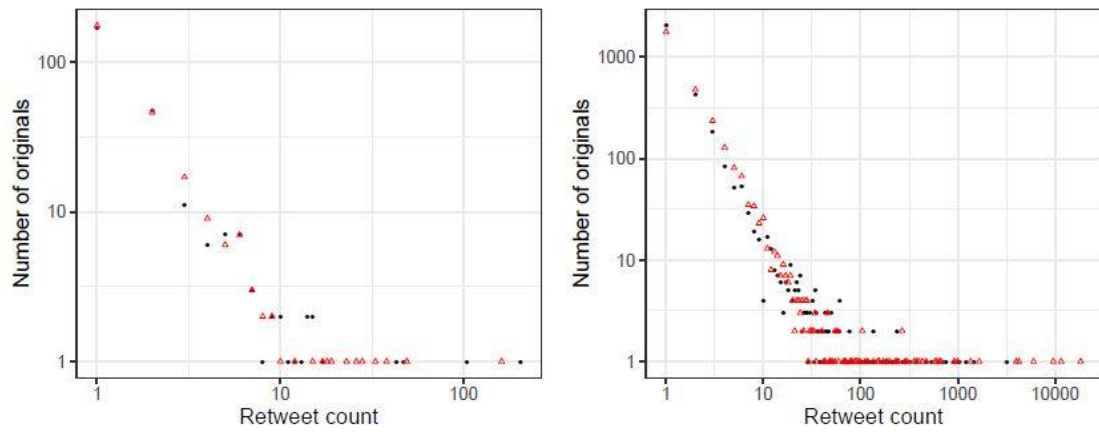
Fig. 12 Frequencies of retweet counts for actual (black, circles) and simulated data (red, triangles) for #thehandmaidstale (left) and #gots7 (right) data.
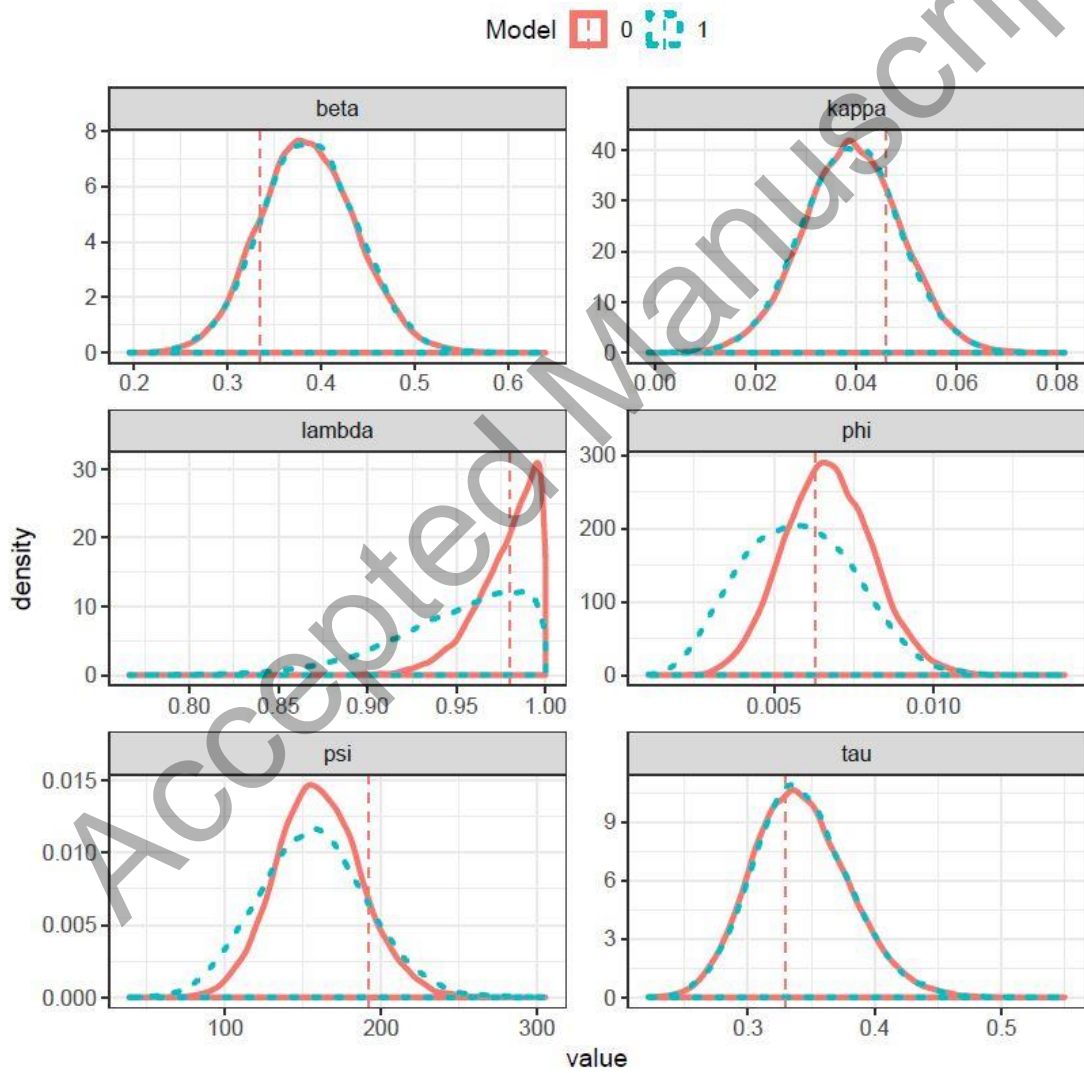

Fig. 13 Posterior density of the parameters (except θ) of models 0 (salmon, solid) and 1 (turqoise, dashed) fitted to the simulated data shown on the left of

Figure 11. The vertical lines represent the respective true values of the parameters.

Table 1 Summaries of originals and retweets for the two data sets.

	#thehandmaidstale	#gots7
Originals	2043	25420
Originals retweeted	265	3145
Total retweets	971	29751
Retweets of top original	204	3204

Table 2 Summaries of chains and computation times (in hours) of the MCMC algorithms for the two data sets. The numbers of originals n and total retweets m are the same as in Table 1.

	#thehandmaidstale	#gots7
Originals (n)	2043	25420
Total retweets (m)	971	29751
Burn-in	1000000	1000000
Thinning	2000	1000
Length of thinned chain	20000	20000
Total number of iterations	41000000	21000000
Time for model 0	28.2	221.5
Time for model 1	69	446.8
Time for GVS	67.6	464.8
Time for RJMCMC	34.7	239.9