Advances in Statistical Analysis manuscript No.
(will be inserted by the editor)

Predicting the Popularity of Tweets Using Internal and External Knowledge: An Empirical Bayes Type Approach

Wai Hong Tan · Feng Chen

Received: date / Accepted: date

Abstract The problem of tweet popularity prediction, or forecasting the total number of retweets stemming from an ancestral tweet, has attracted considerable interest recently. The prediction can be accomplished by fitting a point process model to the sequence of retweet times up to a certain censoring time and project the fitted model to a future time point. However, models employing such approach tend to have inferior prediction accuracy when the censoring time is too short before sufficient information can accumulate. To overcome this, we propose an empirical Bayes type approach of parameter estimation to combine internal knowledge on the times of historical retweets up to the censoring time and external knowledge on complete retweet sequences in the training data. We demonstrate the approach using several point process models with finite-dimensional parameters, where the prior distribution for the parameter of each model is constructed based on the external knowledge, and the likelihood is calculated based on the internal knowledge. The mode of the posterior distribution is used as the estimator of the finite-dimensional parameter, and the mean of the predictive distribution for the number of retweets implied by each of the estimated models is used to predict the tweet popularity. Using a large Twitter data set, we reveal that the proposed methodology not only enables prediction at time zero before the arrival of any retweet event, but also substantially improves the prediction performances of existing models, especially at earlier censoring times.

Keywords empirical Bayes · kernel smoothing · maximum a posteriori (MAP) estimation · nonparametric regression

Wai Hong Tan
UNSW Sydney, Universiti Malaysia Kelantan
E-mail: waihong.tan@unswalumni.com
wai.hong@umk.edu.my

Feng Chen
UNSW Sydney
E-mail: feng.chen@unsw.edu.au

1 Introduction

Twitter is a well-known microblogging platform. Users of Twitter can publish short posts called tweets. A tweet published by a user can be shared by other users as retweets. The retweets themselves can be further retweeted, leading to a cascade or sequence of retweets stemming from the original tweet.

Tweet popularity prediction is an emerging area of research that has attracted considerable interest recently. It has been used primarily to resolve the problems of information overload on Twitter, although it also has other usage scenarios, such as approximating the citation counts of research articles (Eysenbach, 2011), assisting marketing firms to maximize revenues through optimal placements of advertisements (Yang and Leskovec, 2011), and serving as a proxy to the popularity of political candidates in election campaigns (Van Aelst et al., 2017).

One specific prediction problem is to predict the number of retweets received by a tweet. This can be viewed as a regression problem, and approached using machine learning techniques such as logistic regression (Hong et al., 2011), support vector regression (Bandari et al., 2012), Naïve Bayes (Ma et al., 2013), deep neural network (Yang et al., 2014), and random forest (Mishra et al., 2016). These methods typically require the features of the tweet content as input, which can be expensive to extract.

Another approach to the prediction problem is to fit a mathematical model to the observed retweet sequence up to a censoring time and project the fitted model to a future time point to make predictions. Examples of such models include the SEISMIC (Self-Exciting Model of Information Cascades; Zhao et al., 2015), the TiDeH (Time-Dependent Hawkes Process; Kobayashi and Lambiotte, 2016), and the MaSEPTiDE (Marked Self-Exciting Process with Time-Dependent Excitation Function; Chen and Tan, 2018).

As most tweets have relatively short lifespans, the ability to make accurate popularity predictions early is desirable. However, the model-based approach does not work when tweet popularity prediction is needed at the time of its publication or shortly after, as the observed retweet sequence at such times is either empty or too short to allow the reliable fitting of a model. Works that address the problem of tweet popularity prediction at or within a short time of the publication of the tweet have been scarce. Although the feature-driven regression methods can be used for this purpose, they typically do not take advantage of the point process nature of the observed retweet history, while exploiting the history can lead to very good popularity predictions.

Combining the feature-based and model-based approaches should deliver better prediction performances than each approach can separately achieve, as demonstrated by the hybrid method of Mishra et al. (2016). Their method works by training a feature-based regression model without using the retweet sequence, fitting a point process model to the observed retweet sequence by the censoring time, and then retraining the regression model with the parameters of the fitted point process as additional features. Despite its improved prediction accuracy, the method does not always work as the step to fit the point process model will fail if the observation time for the retweet process is too short, making unavailable the extra features required in the final step.

To overcome the aforesaid issues, we propose an alternative approach to combine the strengths of the feature-based and the model-based approaches. Similar to the model-based approach, we fit a point process model to the retweet sequence and project the fitted model to future times for predictions. However, when fitting the model, instead of the maximum likelihood (ML) method used by Chen and Tan (2018) or the least squares (LS) method used by Kobayashi and Lambiotte (2016), we adopt a Bayesian approach and use the *maximum a posteriori* (MAP) method to estimate the model parameters, where the prior distribution for the parameters is constructed from the training data using a feature-based regression method, and the likelihood function is still based on the observations of the retweet sequence up to the censoring time, as in the ML method or the LS method. As the prior distribution is empirically motivated, we term our approach an *empirical Bayes (EB)* type approach.

At censoring time zero, no retweets have been observed and the likelihood function equals 1. Therefore, the posterior of the model parameters is the same as the prior, implying that the MAP estimator is simply the mode of the prior. This means that at the time of publication of a tweet, we already have an estimate of the point process model for its retweets, which can be used for popularity prediction. At later censoring times, the likelihood function weighs increasingly heavily in determining the posterior distribution and the MAP estimator gradually shifts towards the ML estimator. The incorporation of external knowledge in the estimation process through the prior distribution not only avoids the failure to produce predictions at early censoring times faced by the model-based method and Mishra et al.'s hybrid method, but also leads to an objective function with better curvature than the (log-)likelihood function at all censoring times. Therefore, our approach generally gives more stable parameter estimates and overall more accurate popularity predictions.

In Section 2, we describe the Twitter data that motivated our work, which contains the posting or publication time of each tweet relative to the beginning of the day and the number of followers of the publisher, together with the times of all the retweets within seven days and the respective numbers of followers of the retweeters. In Section 3, we show how the proposed EB type approach can be employed on different point process models. In Section 4 we present the results of applying the proposed methodology on the Twitter data and compare the prediction performances of various state-of-the-art approaches in the literature. Finally, Section 5 concludes with a discussion.

2 The Twitter data

Throughout this work, we shall use the Twitter data[1] collected by Zhao et al. (2015) to demonstrate our modeling and prediction methodologies. The data

[1] http://snap.stanford.edu/seismic/

consists of 166,069 reasonably popular tweets, each with at least 49 retweets within seven days of its publication, collected over a 15-day period. Following Zhao et al. (2015) and Chen and Tan (2018), we use the 71,815 tweets collected in the first seven days and the 94,254 tweets collected in the next eight days as the training and test data sets respectively.

For each tweet and retweet, the publication time in days relative to the beginning of the first day of data collection and the number of followers of the corresponding tweeter or retweeter are available. For each retweet, it is known which ancestral (original) tweet it refers to, but it is not known if it directly retweets the ancestral tweet or a previous retweet, which implies that the network structure of the retweets is not available. The contents of tweets were also not included in the data, thus tweet features other than the posting times are not available. See Fig. 1 for a plot of a random retweet sequence.

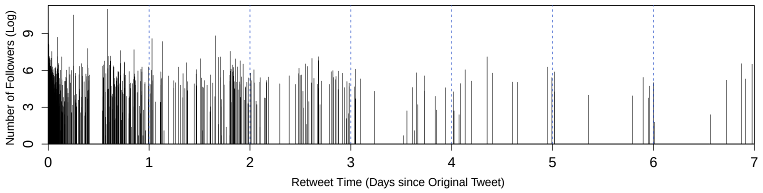

Fig. 1 Plot of a randomly selected retweet sequence, where each bar represents a retweeter with the corresponding followers plotted on the vertical axis (on the logarithmic scale).

The total numbers of retweets accumulated at the end of the observation period are highly skewed, ranging from 49 to 33,484 with mean 205.5 and median 109 in the training data, and from 49 to 17,183 with mean 210.7 and median 110 in the test data. The empirical cumulative distribution of the retweet times relative to the original tweet publication times is shown in Table 1. It can be observed that approximately half of the total numbers of retweets have happened within one hour since the publications of the original tweets, thereby exhibiting the transient nature of tweets. The right-skewed distribution of the retweet times seems compatible with the heavy-tailed distributions of human response times found in the study of other activities, such as e-mail correspondence (Malmgren et al., 2008). Such observation motivates us to use a heavy-tailed function, for example the power-law function, when modeling the variation of the retweet intensity over time.

Table 1 The percentages of retweets that occurred up to each censoring time in the training data. The majority of retweets have happened in the first 12 hours.

Censoring time (hours)	1	2	3	4	5	6	12	168
% of retweets	51.1	59.1	63.8	67.1	69.6	71.6	79.0	100.0

3 The models, the estimation and the prediction methodologies

In all the models considered in this work, we denote the publication time of the original tweet by t^0, the number of followers of the original publisher of tweet by n^0, the retweet times relative to t^0 by $\tau_1 < \tau_2 < \dots$, and the corresponding numbers of followers of the retweeters by $n_1, n_2, \dots$, respectively. The superscript 0 in n^0 and t^0 is used as a reminder that these quantities are available at the time of publication of the original tweet. Let $N(t) = \#\{i \geq 1 : \tau_i \leq t\}$ count the number of retweets up to time $t \geq 0$ relative to t^0. Then $N(t),\ t \geq 0$ is a counting process or a point process. We denote its (conditional) intensity process by $\lambda(t),\ t \geq 0$, so that

$$\lambda(t) = \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \mathbb{E}\left[N(t + \Delta t) - N(t) | \tau_j, n_j, j = 1, \dots, N(t-)\right],$$

with $N(t-)$ denoting the number of retweets right before time t. For a small increment in time $\Delta t > 0$, $\mathbb{E}\left[N(t + \Delta t) - N(t) | \tau_j, n_j, j = 1, \dots, N(t-)\right] \approx \lambda(t)\Delta t$, so $\lambda(t)$ represents the number of retweets per unit time to be expected at time t, given the history of the retweet process prior to time t. In other words, $\lambda(t)$ is the instantaneous event rate, which indicates how fast the retweets are appearing at time t.

If the censoring time T is such that enough retweets have been observed by T, then the ML method can be used to fit the intensity model, and the fitted intensity can be used to predict the number of events to be expected until a future time point. However, if the censoring time is too small and so is the number of retweets $N(T)$, then the optimization of the log-likelihood function can be numerically unstable, and the MLE might not even be well-defined. This issue of the ML method is largely caused by the fact that it relies solely on knowledge internal to the tweet whose popularity is to be predicted, or more specifically, its retweet history by the censoring time, while the external knowledge of the many retweet sequences originating from other tweets in the training data set is ignored altogether. To address this issue, we propose an empirical Bayes type approach to incorporate both internal and external knowledge in the parameter estimation process, where the external knowledge is used to construct the prior distribution for the parameters while the internal knowledge is still used to define the likelihood.

Although our empirical Bayes type approach can be used on any point process model, we shall first demonstrate its use on a relatively simple Poisson process model, before showing how to apply it on more sophisticated models.

3.1 An inhomogeneous Poisson process model

With the Poisson process model, the sequence of retweet times is modelled by an inhomogeneous Poisson process with a time-dependent intensity function $\lambda(t)$, which is assumed to take the following form,

$$\lambda(t) = p(t)d(t), \tag{1}$$

where $p(t)$ reflects the effect of the age of the original tweet on its retweet intensity at time t, hereinafter referred to as the *infectivity function*, and $d(t)$ reflects the time-of-day effect. As the older a tweet gets, the less likely it will get retweeted, the function $p(\cdot)$ should be decreasing. Following Malmgren et al. (2008), we assume it decreases at polynomial rate, so that

$$p(t) = \alpha(1 + \beta t)^{-\gamma},$$

for parameters $\alpha > 0$, $\beta > 0$, and $\gamma > 0$. The parameter α indicates how high the intensity is initially at time $t = 0$, β indicates how soon the intensity decays to half its initial value, and γ indicates how fast the intensity decays over time. They are referred to as the magnitude, the scale, and the shape parameters respectively. These parameters are assumed to be tweet-specific, and may differ for retweet sequences originating from different tweets.

The nonnegative function $d(\cdot)$ is introduced to account for the possible time-of-day effect related to the circadian rhythm of human activity, and is assumed to be periodic with period one day. Specifically, if we measure time in days, then there is a function $\rho(\cdot) \geq 0$ such that

$$d(t) = \rho(t^0 + t - \lfloor t^0 + t \rfloor), \tag{2}$$

where $\lfloor x \rfloor$ indicates the greatest integer $\leq x$. Note that the argument t in (2) refers to the time since the publication time t^0 of the original tweet, and t^0 is measured in (fractional) days relative to the beginning of the day on which it was published. Thus, $t^0 + t - \lfloor t^0 + t \rfloor \in [0, 1)$ refers to the time, in fractions of a day, relative to the beginning of the day on which $t^0 + t$ falls.

In addition to the periodicity assumption, the function $d(\cdot)$ in (2) is assumed to be smooth. For identifiability, we also assume that the function $\rho(\cdot)$ integrates to unity, so that it is a probability density function supported by $[0, 1)$. The smoothness and periodicity of the function $d(\cdot)$ imply that the function $\rho(\cdot)$ is smooth, and furthermore satisfies the continuity condition,

$$\rho(0) = \lim_{t \downarrow 0} \rho(t) = \lim_{t \uparrow 1} \rho(t). \tag{3}$$

For convenience, the function $\rho(\cdot)$ shall be referred to as the *rhythm function*.

3.1.1 An empirical Bayes type approach to parameter estimation

The empirical Bayes type approach we propose is simply a Bayesian approach with the prior distribution empirically constructed from the training data. To discuss the construction of the prior distribution for the parameters $\theta = (\alpha, \beta, \gamma)$ of the Poisson model, we first discuss their ML estimation. The log-likelihood of θ relative to a retweet sequence up to a censoring time T is given by Daley and Vere-Jones (2003),

$$\ell(\theta) = \sum_{i=1}^{N(T)} \log \lambda(\tau_i; \theta) - \int_0^T \lambda(t; \theta)\, \mathrm{d}t. \tag{4}$$

To obtain the ML estimator of θ, we maximize (4), with $\lambda(t)$ set to $\lambda(t;\theta) = p(t;\theta)d(t) = \alpha(1+\beta t)^{-\gamma}d(t)$, as a function of θ. Note however that the log-likelihood depends on the unknown rhythm function $\rho(\cdot)$ discussed above via the function $d(\cdot)$. The function $\rho(\cdot)$ is unspecified except for the smoothness condition. For simplicity, we shall fix it at an estimated value $\hat{\rho}(\cdot)$ and treat it as known when estimating the other parameters θ.

Nonparametric estimation of the rhythm function
Since we have assumed that the rhythm function $\rho(\cdot)$ is a density function representing the distribution of the tweet publication times in a day, we estimate $\rho(\cdot)$ nonparametrically using the kernel density estimator (KDE; see Silverman, 1986, Section 2.4) based on the publication times of the original tweets in the training data set. To correct for the boundary effects suffered by the KDE, and to ensure the continuity condition in (3), we use a pseudodata approach similar to the data reflection approach discussed in Section 2.10 of Silverman (1986) and the pseudodata approach of Cowling and Hall (1996).

Specifically, if $t_1^0, \ldots, t_n^0 \in [0,1)$ denote the data, that is, the tweet publication times measured in days since 00:00:00 on the days they were posted, we augment the data by adding $t_1^0 - 1, \ldots, t_n^0 - 1$ and $t_1^0 + 1, \ldots, t_n^0 + 1$. Following this, we estimate the density on $[0,1)$ using the KDE with the augmented data, and subsequently rescale the estimates so that the estimated density curve $\hat{\rho}(\cdot)$ integrates to unity. Finally, the estimated time-of-day effect function $d(\cdot)$ for a retweet sequence originating from a tweet published at time t^0 is

$$\hat{d}(t) = \hat{\rho}(t^0 + t - \lfloor t^0 + t \rfloor). \tag{5}$$

In our implementation of the KDE, we have used the function `density` from the `stats` package of R (R Core Team, 2019), with the biweight kernel $K(x) = 15/16(1-x^2)_+^2$ and the bandwidth selected using the default normal reference distribution approach based on the unaugmented data.

The prior distribution for model parameters
To construct the prior distribution for the parameters of the Poisson model for a specific retweet sequence, we first compute the ML estimate for each of the complete retweet sequences in the training data set, and denote these estimates by $\hat{\theta}_i^0 = (\hat{\alpha}_i^0, \hat{\beta}_i^0, \hat{\gamma}_i^0)$, $i = 1, \ldots, 71815$. Recall that the complete retweet sequences in the training data set are all fairly long, so the numerical stability of the ML estimator is not an issue here.

Next, we fit three separate nonparametric regression models with $y_i = \log \hat{\alpha}_i^0$, $\log \hat{\beta}_i^0$ and $\log \hat{\gamma}_i^0$ as the respective response variables, and $x_i = (\log(n_i^0 + 1), t_i^0) = (m_i^0, t_i^0)$ as the input variables or features, using the locally weighted scatter-plot smoother (LOESS; Cleveland and Devlin, 1988). In our numerical implementation of the regression, we have used the `loess` function from the `stats` package of R, with the degree of the local polynomial set to 2, the kernel function set to the default tricubic kernel $K(x) = \frac{70}{81}(1-|x|^3)_+^3$, and the respective span parameters selected using the generalized cross validation (GCV; Golub et al., 1979) method.

Then, for a tweet posted at time t^0 by a tweeter with n^0 followers (hence with features $x = (m^0, t^0)$), we predict the three components of the log-parameter vector $\eta = (\eta_1, \eta_2, \eta_3) \equiv (\log\alpha, \log\beta, \log\gamma)$ separately using the respective nonparametric models obtained in the last step, and denote the predicted log-parameter vector by $\tilde{\eta}^0 = (\tilde{\eta}_1^0, \tilde{\eta}_2^0, \tilde{\eta}_3^0) \equiv (\log\tilde{\alpha}^0, \log\tilde{\beta}^0, \log\tilde{\gamma}^0)$ and the standard errors by $(\tilde{e}_1, \tilde{e}_2, \tilde{e}_3)$. For implementation, we have used the `predict.loess` function in `R`.

Finally, we define the prior density function for the log-parameter vector η as follows,

$$\pi(\eta) = f(\eta_1; \tilde{\eta}_1^0, \tilde{e}_1^2) f(\eta_2; \tilde{\eta}_2^0, \tilde{e}_2^2) f(\eta_3; \tilde{\eta}_3^0, \tilde{e}_3^2),$$

with $f(\cdot; \mu, \sigma^2)$ denoting the normal density function with mean μ and variance σ^2, so that $\tilde{\eta}^0$ comes as the maximizer of $\pi(\eta)$. Here, we note that the prior distributions for the log-parameters η_1, η_2, and η_3 are the respective *confidence distributions* (Xie and Singh, 2013) based on the training data for their means $\mathbb{E}[\eta_i]$, $i = 1, 2, 3$, when they are treated as random variables with means depending on the features $x = (m^0, t^0)$.

The empirical Bayes estimator

The posterior density function for the log-parameters, up to a normalizing constant, is $\pi(\eta)\exp(\ell(e^\eta))$. The maximizer of the posterior density, or equivalently, the maximizer of its logarithm (up to an additive constant)

$$\tilde{\ell}(\eta) = \log\pi(\eta) + \ell(e^\eta), \tag{6}$$

is a Bayes estimator of η, called the *maximum a posteriori (MAP)* estimator. As a reminder, the $\ell(\theta)$ here is as in (4) with the $\lambda(\cdot)$ given by (1) and the function $d(\cdot)$ replaced by its estimator (5). We denote the MAP estimator by $\tilde{\eta}$, and define our EB estimator at time T for the parameters θ as $\tilde{\theta} = e^{\tilde{\eta}}$.

As mentioned earlier, at censoring time zero, the log-likelihood in (4) is 0, and so the maximizer of the prior density function, that is, $\tilde{\eta}^0$, is also the maximizer of the posterior density, and therefore $e^{\tilde{\eta}^0}$ will be taken as the estimator of the tweet-specific model parameters, enabling tweet popularity prediction. For reference, Figure A.1 summarizes the steps involved in obtaining the EB estimates of the parameters.

3.1.2 Predicting the popularity

After the parameters are estimated, the model for a specific retweet time sequence is identified. We can then use the mean or median of the predictive distribution of the number of retweets implied by the identified model from the censoring time T to a future time $\tilde{T}$, plus the number of retweets observed by time T, as a point prediction of the total number of retweets by $\tilde{T}$.

For the Twitter data considered in this work, since we know a priori that the final popularity value is at least 49, the mean and median of the distribution for the number of future retweets should be calculated conditional on $N(\tilde{T}) - N(T) \geq 49 - N(T)$. Under the Poisson process model, $N(\tilde{T}) - N(T)$

is Poisson distributed with its mean equal to the integral of the identified intensity function from T to $\tilde{T}$, $\int_T^{\tilde{T}} \lambda(t;\tilde{\theta})\,\mathrm{d}t = \int_0^{\tilde{T}-T} \lambda(T+t;\tilde{\theta})\,\mathrm{d}t$. Thus, the computations of its conditional mean and median are straightforward.

To compare the performances of different prediction models, we shall use the mean and median absolute percentage errors (MAPE and MdAPE), which have been used by various works in the literature (Zhao et al., 2015; Kobayashi and Lambiotte, 2016; Chen and Tan, 2018). For convenience, the Poisson model proposed here with its parameters estimated using the EB type approach shall be referred to as the EB Poisson model. When the EB approach is applied on the MaSEPTiDE model and the TiDeH model, discussed below in Sections 3.2 and 3.3, these models shall be similarly termed the EB MaSEPTiDE model and the EB TiDeH model respectively.

3.2 The MaSEPTiDE model

The MaSEPTiDE model is a (marked) point process model that has a self-exciting feature similar to how a tweet excites retweets and retweets excite further retweets, and has been shown to produce more accurate tweet popularity predictions than competing models (Chen and Tan, 2018). Under this model, the intensity process $\lambda(t)$ of $N(t)$ is given by

$$\begin{aligned}
\lambda(t) &= \alpha\phi(t;\delta_1,\delta_2) + \sum_{i=1}^{N(t-)} p(\tau_i;\beta) r(n_i;\gamma)\phi(t-\tau_i;\delta_1,\delta_2),\\
p(\tau;\beta) &= e^{-\beta\tau},\\
r(n;\gamma) &= \gamma\log(n+1),\\
\phi(t;\delta_1,\delta_2) &= \frac{\delta_2(\delta_1-1)}{\delta_1}\left(1+\frac{\delta_2 t}{\delta_1}\right)^{-\delta_1},
\end{aligned} \tag{7}$$

where as before, $N(t-)$ denotes the number of retweets up to, but not including, time t. The retweeters' follower counts n_i's are assumed to be independent of past retweet times and i.i.d. (independent and identically distributed) lognormal with $\mathbb{E}[\log(n_i+1)] = M$ and $\mathbb{V}\mathrm{ar}(\log(n_i+1)) = \sigma^2$. The log-likelihood function of the parameters $\theta = (\alpha,\beta,\gamma,\delta_1,\delta_2,M,\sigma^2)$ relative to the retweet sequence observed up to the censoring time T is similar to (4), although it has an extra term to account for the likelihood contribution of the event marks $m_i = \log(n_i+1)$:

$$\ell(\theta) = \sum_{i=1}^{N(T)} \log\lambda(\tau_i) - \int_0^T \lambda(t) + \log f(m_i;M,\sigma^2), \tag{8}$$

with $\lambda(t)$ given in (7) and $f(\cdot;\mu,\sigma^2)$ denoting the normal density function with mean μ and variance σ^2 as before.

Despite its ability to make relatively accurate popularity predictions at early censoring times, with the ML method of model fitting the MaSEPTiDE model either gives unreasonably large popularity predictions or fails to produce a prediction at all when the censoring time is too small. Fortunately, this issue can be resolved by adopting the EB approach.

As in the Poisson model, the ML estimates of the parameter vector $\theta = (\alpha, \beta, \gamma, \delta_1, \delta_2, M, \sigma^2)$ in the MaSEPTiDE model (7) can be obtained on the complete retweet sequences in the training data, and subsequently used to construct the prior distribution for model parameters used in the EB estimation approach. Specifically, the prior density for the transformed parameters $\eta = (\eta_1, \ldots, \eta_7) \equiv (\log\alpha, \log\beta, \log\gamma, \log\delta_1, \log\delta_2, M, \log\sigma^2)$ of the MaSEPTiDE model for a retweet sequence with features $x = (t^0, m^0)$ is

$$\pi(\eta) = \prod_{k=1}^{7} f(\eta_k; \tilde{\eta}_k^0, \tilde{e}_k^2) \tag{9}$$

where $f(\cdot)$ denotes the normal density function as before, and $\tilde{\eta}_k^0$ and $\tilde{e}_k^0$ denote respectively the LOESS estimate of η_k^0 and the associated standard error. The EB estimate of η can then be obtained as the maximizer of the logarithm of the posterior density function, which is formally given, up to an additive constant, by (6), with $\pi(\cdot)$ given in (9) and $\ell(\cdot)$ in (8).

After the MaSEPTiDE model is estimated, the number of retweets up to a future time point can be predicted using the solve-the-equation method or the simulation-based method described in Section 2.5 of Chen and Tan (2018). The former involves solving an integral equation for the conditional expectation of the number of retweets given the observed retweet history as a function of the future time point, and evaluating the solution function at the desired time point. The latter involves simulating the retweet sequence from time T to $\tilde{T}$ given the retweet history by T according to the fitted model.

In the MaSEPTiDE model, the lognormality of the retweeters' follower counts n_i's is assumed to facilitate the estimation of the parameter M using the EB approach. An alternative approach that does not require the lognormality, or any other parametric assumption on the distribution of n_i's is the following: when no retweet events have been observed for the target tweet, we choose the retweet sequence in the training data set whose ancestral tweet is the closest to the target tweet by the features (m^0, t^0), and use the empirical distribution of the follower counts $n_1, n_2, \ldots, n_{N(\tilde{T})}$ as an estimate of the distribution of the follower counts, and when retweet events have arrived, we revert to using the empirical distribution of the follower counts specific to the target tweet's retweeters. This alternative approach worked fairly well in our numerical experiments.

3.3 The TiDeH model

The TiDeH model (Kobayashi and Lambiotte, 2016) is a special case of the SEISMIC model (Zhao et al., 2015), where the intensity process of $N(t)$ is

$$\begin{aligned}\lambda(t) &= p(t;\alpha,\beta,\gamma,\delta)\sum_{i=0}^{N(t-)} n_i\bar{\phi}(t-\tau_i),\\ p(t;&\alpha,\beta,\gamma,\delta) = \alpha(1-\beta\sin(2\pi(t+\gamma)))e^{-t/\delta}.\end{aligned}$$

The damped sinusoid form of the function $p(\cdot)$ accounts for the repetitiveness of human routine activities and the decreasing likelihood of a tweet being retweeted, with the parameters α,β, γ and δ representing the overall retweet intensity, the relative amplitude of oscillation and its phase, and the characteristic time of popularity decay respectively. The function $\bar{\phi}(\cdot)$ is a power-law kernel with parameters common to all the retweet sequences considered, estimated ad hoc from some long retweet sequences selected from the training data (Zhao et al., 2015) and treated as known when estimating other parameters.

The parameters $\theta = (\alpha,\beta,\gamma,\delta)$ in the infectivity function $p(\cdot)$ can be estimated via a two-step approach based on nonparametric smoothing and least squares (Kobayashi and Lambiotte, 2016) or the ML method (Chen and Tan, 2018). Both methods require some retweet events to be observed before estimation is possible. To overcome this constraint, we can again adopt the EB estimation approach and estimate the parameters using the MAP method. Specifically, we use the following normal prior for the transformed parameters $\eta = (\eta_1,\eta_2,\eta_3,\eta_4) \equiv (\log\alpha,\beta,\gamma,\log\delta)$,

$$\pi(\eta) = \prod_{i=1}^{4} f(\eta_i;\tilde{\eta}_i^0,\tilde{e}_i^2)$$

with $\tilde{\eta}^0 = (\tilde{\eta}_1^0,\tilde{\eta}_2^0,\tilde{\eta}_3^0,\tilde{\eta}_4^0) \equiv (\log\tilde{\alpha}^0,\tilde{\beta}^0,\tilde{\gamma}^0,\log\tilde{\delta}^0)$ and $(\tilde{e}_1,\tilde{e}_2,\tilde{e}_3,\tilde{e}_4)$ being the predicted parameters and their standard errors by the LOESS regression method. Note here that the logarithmic transformation of the model parameters β and γ is not used as they can take negative values.

The future numbers of retweets can be predicted via the solve-the-equation method or the simulation-based method detailed in Chen and Tan (2018). For fair comparison with the EB Poisson model, the a priori knowledge on the lower bound of the popularity should be incorporated into predictions based on the EB MaSEPTiDE and the EB TiDeH models. This can be done by repeatedly simulating the number of retweets from the censoring time T to $\tilde{T} = 7$ days and retaining only those numbers $\geq 49 - N(T)$, and when none of the simulated retweet counts meets the condition, the mean and median of the predictive distribution can simply be approximated by the lower bound.

4 Application to the Twitter data

In this section we present the results of applying the models and prediction methods discussed in Section 3 to the Twitter data discussed in Section 2. In particular, we compare their prediction performances with state-of-the-art methodologies in the literature and with each other.

4.1 The EB Poisson model

The estimated rhythm function $\rho(\cdot)$ is shown in Figure 2, which shows that the peak hours of tweet activity are between 23:00 and 03:00 UTC (Coordinated Universal Time) or 6:00 to 10:00 PM CST (Central Standard Time), with its lowest point hovering around 14:00 UTC or 9:00 AM CST. This suggests that a tweet is more likely to attract retweets between 23:00 and 03:00 UTC.

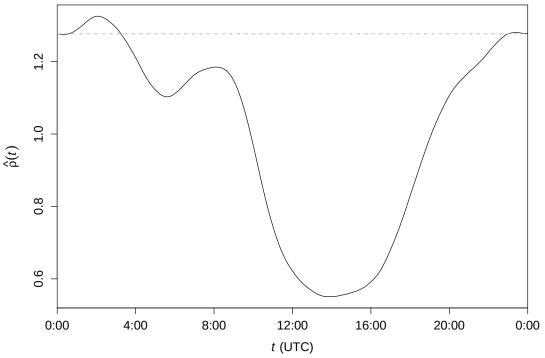

Fig. 2 Estimated function $\rho(\cdot)$ showing the diurnal patterns of Twitter users' activity levels, which suggests that the peak hours of activity are between 23:00 and 03:00 UTC.

The EB estimates of the log-parameters, together with their corresponding ML estimates using only the internal history of the retweet sequence, at different censoring times for two randomly selected retweet sequences in the test data set are illustrated in Figure 3. Note, at time zero, Figure 3 only shows the EB estimates, but not the ML estimates, as the ML estimates are unavailable due to the lack of any observations. The figure reveals that the EB estimates at different times are substantially more stable compared to the ML estimates, suggesting that the use of prior distribution has a strong regularization effect on the ML estimates. Essentially, the EB type approach is a *penalized maximum likelihood* approach whereby a concave quadratic penalty is added to the log-likelihood to penalize parameters far from the initial LOESS estimates of the parameters. The presence of the quadratic penalty causes the penalized log-likelihood to have larger curvature than its unpenalized counterpart, thus leading to more stable estimators of the parameters.

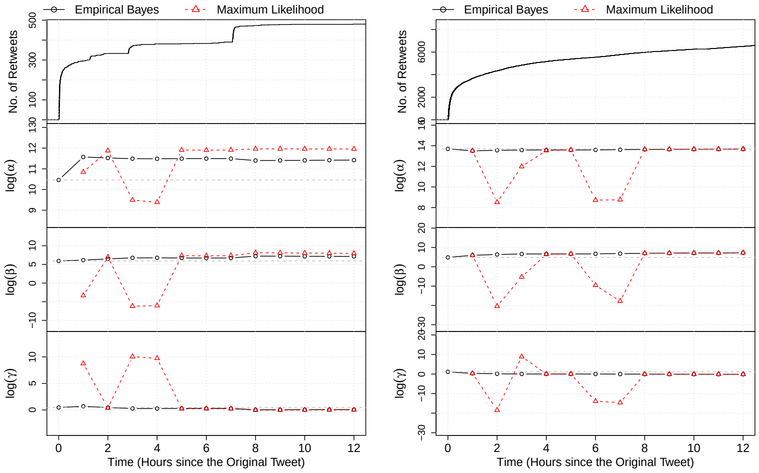

Fig. 3 Estimates of the log-parameters for the Poisson process model using the empirical Bayes (EB) type approach and maximum likelihood (ML) approach at different censoring times, for two random cascades. The top panel of each subfigure shows the sample path of the counting process $N(t)$ for the corresponding retweet sequence up to 12 hours, the lower panels of each subfigure show the estimated log-parameters at $T = 0, 1, \ldots, 12$ hours.

For the two sample cascades in Figure 3, we show in Table 2 the predicted popularity values $N(\tilde{T})_{\text{pred}}$ based on the ML and EB estimates obtained at censoring times $T = 0, 1, 2, 3$ hours, and the associated APE values. The comparisons at later censoring times are similar, and therefore not shown. It can be observed from the APEs in Table 2 that, the predictions obtained based on the EB type approach are consistently more accurate than those based on the ML approach. In addition, the predictions obtained from the ML approach seem very volatile when the tweet in question is highly popular, in contrast to the increasingly more accurate predictions yielded using the EB type approach.

Table 2 The observed popularity, the prediction and APE values for the Poisson model using the ML and EB estimation approaches, and the final popularity at $\tilde{T} = 7$ days for each of the two sample cascades in Figure 3 at censoring times $T = 0, 1, 2, 3$ hours.

	T	$N(T)$	ML		EB		$N(\tilde{T})$
	(hours)		$N(\tilde{T})_{\text{pred}}$	APE	$N(\tilde{T})_{\text{pred}}$	APE	
Sample 1	0	0	-	-	194.73	60.90	498
	1	296	296.03	40.56	310.32	37.69	
	2	333	362.31	27.25	371.37	25.43	
	3	370	371.25	25.45	430.94	13.47	
Sample 2	0	0	-	-	3718.62	61.25	9597
	1	3670	5034.44	47.54	5077.92	47.09	
	2	4338	38535.07	301.53	6558.05	31.67	
	3	4846	4898.07	48.96	7721.68	19.54	

To benchmark the EB Poisson method against the state-of-the-art prediction methodologies, we compare the prediction performance of our model with that of the MaSEPTiDE and the TiDeH models, since they have been shown to outperform pre-existing methods such as the SEISMIC (cf. Kobayashi and Lambiotte, 2016; Chen and Tan, 2018). Figure 4 shows the boxplots of APEs based on the final popularity predictions at $\tilde{T} = 7$ days, by the EB Poisson model, the (ML) Poisson model, the MaSEPTiDE model, and the TiDeH model, at censoring times $T = 0, 1, \ldots, 12$ hours. As the distributions of the prediction APEs of these methods have very long right tails, especially those with the MaSEPTiDE model, the outlying APEs have not been shown for better visualization. Note that at time zero, only the EB Poisson model can produce predictions but not the other models due to the unavailability of the parameter estimates. Also, as the smoothing parameter used in the nonparametric estimation of the TiDeH model infectivity function is set to one hour, the approach cannot produce any meaningful predictions at $T = 1$ hour, and is therefore excluded from comparison at that time.

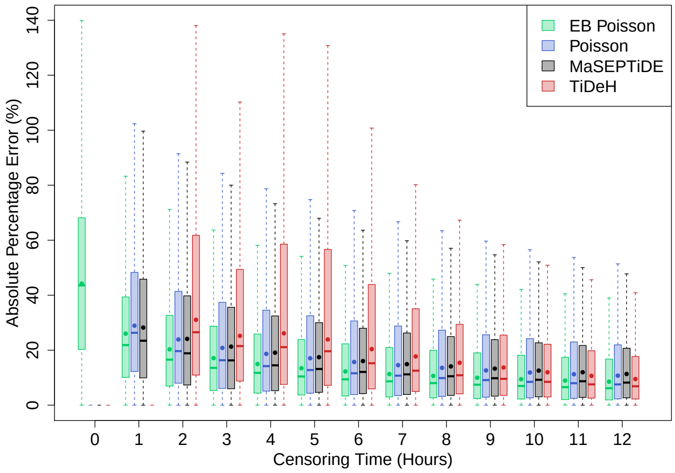

Fig. 4 The APEs of different prediction methods across different censoring times at $T = 0, 1, \ldots, 12$ hours. The Poisson model is also included for comparison. The circular point in each boxplot shows the MAPE, while the horizontal thick bar shows the MdAPE. The EB Poisson model is clearly the best performing model at all the censoring times, and is able to make a prediction even at time zero.

Figure 4 shows that both the MAPE and MdAPE values decrease as the censoring time increases for each method considered, thereby indicating a gradual improvement in the prediction accuracy. It can also be seen that although the Poisson model-based approach (with ML estimates of parameters) is not competitive at all censoring times, the EB Poisson model consistently outperforms the other competing approaches across the censoring times according to both metrics. For fair comparisons, the predictions by all the three models considered have incorporated the extra knowledge on the lower bound of the final popularity value. However, even without incorporating such knowledge, the EB Poisson model would still stand out as the best performing model.

4.2 The EB MaSEPTiDE and the EB TiDeH models

This section presents the popularity prediction results by the EB MaSEPTiDE and the EB TiDeH models, and compare them with the EB Poisson model. We calculated the MAPEs and MdAPEs of predictions by the three models with observations up to censoring times $T = 0, 1, 2, \ldots, 12$ hours, and discovered that the EB Poisson model outperforms both the EB MaSEPTiDE and the EB TiDeH models at time zero, but the EB MaSEPTiDE model outperforms both the EB Poisson and the EB TiDeH models from $T = 1$ hour onward.

For fine-grained comparisons between the EB Poisson and the EB MaSEPTiDE models, we show in Table 3 the prediction error metrics by these two models at various censoring times between $T = 0$ and $T = 60$ minutes, which shows that the EB MaSEPTiDE model outperforms the EB Poisson model (and the EB TiDeH model) from $T = 3$ minutes onward. These results suggest that at early censoring times when fewer retweets are observed, simpler models such as the Poisson process model tends to achieve more accurate popularity predictions, while at later censoring times when more events accumulate, sophisticated models such as the MaSEPTiDE model is better at capturing the dynamics among the retweet events, thereby producing more accurate predictions. On another note, upon inspecting the goodness-of-fit (GOF), the EB models, which generally predict better than their ML counterparts, pass the GOF tests at various significance levels on noticeably lower percentages of cascades. This implies that models with better fit to historical data do not necessarily have better out-of-sample prediction performances.

Table 3 The prediction MAPEs and MdAPEs of the EB Poisson and the EB MaSEPTiDE models, at various censoring times between $T = 0$ and $T = 1$ hour. The EB MaSEPTiDE model outperforms the EB Poisson model from $T = 3$ minutes onward.

T	MAPE (%) for EB		MdAPE (%) for EB	
(minutes)	Poisson	MaSEPTiDE	Poisson	MaSEPTiDE
0	47.9	196.2	43.6	62.0
1	44.6	66.4	37.7	42.5
2	50.6	55.4	37.9	39.6
3	48.8	47.7	37.3	36.8
4	44.8	43.4	35.3	34.7
5	41.5	40.4	33.4	32.9
10	35.2	33.5	29.6	27.9
20	33.0	29.4	28.3	24.3
30	31.2	27.7	26.5	22.2
40	29.8	26.4	24.6	20.7
50	28.6	25.3	23.2	19.4
60	27.4	24.4	21.8	18.3

5 Conclusion and discussion

In this paper, we have proposed an empirical Bayes (EB) type approach, which uses the maximum a posteriori (MAP) method to estimate the parameters of point process models based on knowledge internal and external to a retweet sequence. With the MAP estimators, the fitted models were found to produce reasonable estimates of tweet-specific parameters, and sensible predictions for all instances of tweets considered.

When comparing the performances of different prediction models with the MAPE and MdAPE as the performance evaluation metrics, we have used the predictive mean as the functional for its ease-of-computation. Employing the EB type approach on a simple Poisson model leads to more accurate popularity predictions, compared to two state-of-the-art popularity prediction methods based on the TiDeH (Kobayashi and Lambiotte, 2016) and the MaSEPTiDE (Chen and Tan, 2018) models respectively. However, when fitted using the EB type approach, these two models have markedly better prediction performances, with the EB MaSEPTiDE model surpasses even the EB Poisson model if the retweet sequence is observed for three minutes or longer.

In constructing the prior distribution for the parameters, although we have used the LOESS regression, other machine learning regression methods can also, in principle, be used for this purpose, as along as the standard error of the regression estimator is available. Comparing the performances of the EB type approach with different methods of prior construction could be an interesting problem for future research.

As a remark, the Twitter data used in this work does not contain other features of the original tweet and its author that are available at or before its publication which might be useful for popularity prediction, such as the account age, the author's geolocation, the length of the original tweet, the presence/absence of external links, images, videos, specific keywords or certain hashtags, or the machine-learned topics of the original tweet (Blei et al., 2003; Kant et al., 2020). When such extra information is available, the EB type approach still applies, and the resulting prediction performance is expected to improve even further, although the LOESS regression step in prior distribution construction may be computationally more expensive.

Acknowledgements The authors gratefully acknowledge the constructive comments from the reviewers, which have led to improved presentation. This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney. The research also benefited from the assistance of resources from the National Computational Infrastructure (NCI), supported by the Australian Government.

References

Bandari R, Asur S, Huberman B (2012) The pulse of news in social media: Forecasting popularity. ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993–1022

Chen F, Tan WH (2018) Marked self-exciting point process modelling of information diffusion on Twitter. Ann Appl Statist 12(4):2175–2196

Cleveland WS, Devlin SJ (1988) Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83(403):596–610

Cowling A, Hall P (1996) On pseudodata methods for removing boundary effects in kernel density estimation. Journal of the Royal Statistical Society Series B (Methodological) 58(3):551–563

Daley DJ, Vere-Jones D (2003) An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods, 2nd edn. Springer-Verlag, New York

Eysenbach G (2011) Can tweets predict citations? metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. Journal of medical Internet research 13(4)

Golub GH, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21(2):215–223

Hong L, Dan O, Davison BD (2011) Predicting popular messages in Twitter. In: Proceedings of the 20th international conference companion on World wide web, ACM, pp 57–58

Kant G, Weisser C, Säfken B (2020) TTLocVis: A Twitter topic location visualization package. Journal of Open Source Software 5(25)

Kobayashi R, Lambiotte R (2016) TiDeH: Time-dependent Hawkes process for predicting retweet dynamics. In: Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016), pp 191–200

Ma Z, Sun A, Cong G (2013) On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology 64(7):1399–1410

Malmgren RD, Stouffer DB, Motter AE, Amaral LA (2008) A Poissonian explanation for heavy tails in e-mail communication. Proceedings of the National Academy of Sciences 105(47):18153–18158

Mishra S, Rizoiu MA, Xie L (2016) Feature driven and point process approaches for popularity prediction. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, ACM, pp 1069–1078

R Core Team (2019) R: A language and environment for statistical computing

Silverman B (1986) Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis

Van Aelst P, van Erkel P, D'heer E, Harder RA (2017) Who is leading the campaign charts? comparing individual popularity on old and new media. Information, Communication & Society 20(5):715–732

Xie M, Singh K (2013) Confidence distribution, the frequentist distribution estimator of a parameter: A review. International Statistical Review 81(1):3 – 39

Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: Proceedings of the fourth ACM international conference on Web search and data mining, ACM, pp 177–186

Yang M, Chen K, Miao Z, Yang X (2014) Cost-effective user monitoring for popularity prediction of online user-generated content. In: 2014 IEEE International Conference on Data Mining Workshop, pp 944–951

Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J (2015) SEISMIC: A self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp 1513–1522

Appendix

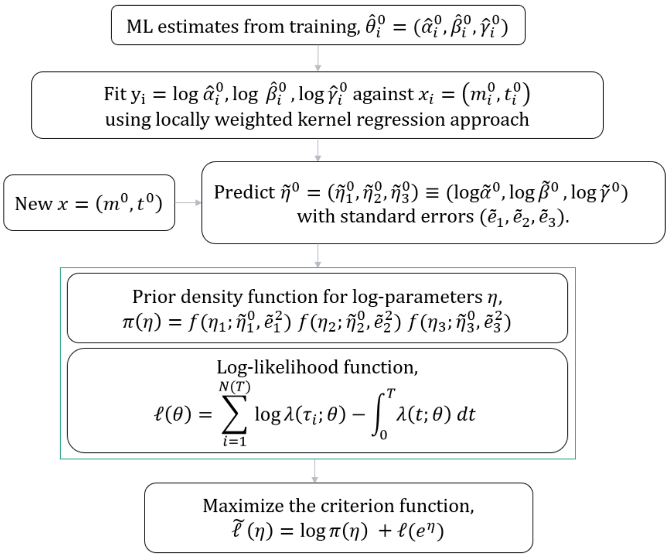
ML estimates from training, $\hat{\theta}_i^0 = (\hat{\alpha}_i^0, \hat{\beta}_i^0, \hat{\gamma}_i^0)$
Fit $y_i = \log \hat{\alpha}_i^0, \log \hat{\beta}_i^0, \log \hat{\gamma}_i^0$ against $x_i = (m_i^0, t_i^0)$ using locally weighted kernel regression approach
New $x = (m^0, t^0)$
Predict $\tilde{\eta}^0 = (\tilde{\eta}_1^0, \tilde{\eta}_2^0, \tilde{\eta}_3^0) \equiv (\log \tilde{\alpha}^0, \log \tilde{\beta}^0, \log \tilde{\gamma}^0)$ with standard errors $(\tilde{e}_1, \tilde{e}_2, \tilde{e}_3)$.
Prior density function for log-parameters η, $\pi(\eta) = f(\eta_1; \tilde{\eta}_1^0, \tilde{e}_1^2)\, f(\eta_2; \tilde{\eta}_2^0, \tilde{e}_2^2)\, f(\eta_3; \tilde{\eta}_3^0, \tilde{e}_3^2)$
Log-likelihood function, $\ell(\theta) = \sum_{i=1}^{N(T)} \log \lambda(\tau_i; \theta) - \int_0^T \lambda(t; \theta)\, dt$
Maximize the criterion function, $\tilde{\ell}(\eta) = \log \pi(\eta) + \ell(e^\eta)$

Fig. A.1 A summary of the procedures involved to obtain the empirical Bayes estimates. The final criterion function combines the knowledge internal and external to a retweet sequence, depending respectively on the current log-likelihood function and the log-prior density function. When the censoring time is at zero, the maximizer of the prior density function is $\tilde{\eta}^0$, and $e^{\tilde{\eta}^0}$ will be taken as the estimator of the tweet-specific model parameters.

Declarations

Funding:
Tan was supported by UMK Fundamental Research Grant [R/FUND/A0100/01348A/001/2020/00840]
Chen was partly supported by UNSW Science Faculty Research Grant [PS35307]

Conflicts of interest:
Not applicable

Availability of data and material:
Data is available from http://snap.stanford.edu/seismic/

Code availability:
Code is available from the authors upon request