MARKED SELF-EXCITING POINT PROCESS MODELLING OF INFORMATION DIFFUSION ON TWITTER

BY FENG CHEN AND WAI HONG TAN

UNSW Sydney

Information diffusion occurs on microblogging platforms like Twitter as retweet cascades. When a tweet is posted, it may be retweeted and henceforth further retweeted, and the retweeting process continues iteratively and indefinitely. A natural measure of the popularity of a tweet is the number of retweets it generates. Accurate predictions of tweet popularity can assist Twitter to rank contents more effectively and facilitate the assessment of potential for marketing and campaigning strategies. In this paper, we propose a model called the Marked Self-Exciting Process with Time-Dependent Excitation Function, or MaSEPTiDE for short, to model the retweeting dynamics and to predict the tweet popularity. Our model does not require expensive feature engineering but is capable of leveraging the observed dynamics to accurately predict the future evolution of retweet cascades. We apply our proposed methodology on a large amount of Twitter data and report substantial improvement in prediction performance over existing approaches in the literature.

1. Introduction. The advancement of technology has dramatically changed the ways people connect to each other over the past few years. This contributes to the increasing popularity of microblogging platforms, which integrate the features of instant messaging and blogging, enabling users to conveniently share contents like short sentences, images or videos. Twitter is a microblogging service that allows the users to share information in the form of 140-character messages called *tweets*. As a tweet is posted by a user, it may be shared by the *followers* of the original poster through an action known as *retweeting*, which involves the publication of a tweet that explicitly refers to the original tweet via its unique identification number, or a *retweet*. This retweeting process can iterate indefinitely, resulting in a cascade of retweets.

Information diffusion modelling in Twitter has been an active field of research. Tumasjan et al. (2010) considered Twitter as a platform used for political deliberation and analyzed tweet sentiments by machine learning to forecast the results of elections. By learning the features, one can also find the likelihood of retweets based on the interestingness of contents [Naveed et al. (2011)]. Matsubara et al. (2012) proposed a model to account for the rise and fall of influence propagation

whereas Alves, Assunção and de Melo (2016) recently modelled random series of events prevalent in Twitter by Poissonian and self-feeding processes. These studies, however, do not emphasize on popularity prediction, which is our primary concern.

The *popularity* of a tweet is naturally measured by the size of the retweet cascade, or the number of retweets it generates. The predictions of tweet popularity are important as they can assist Twitter to rank contents more effectively and facilitate the assessment of potential for marketing and campaigning strategies. As such, many models have been proposed to capture the retweeting dynamics and to predict the popularity. One noticeable work is by Zaman, Fox and Bradlow (2014), who proposed a Bayesian approach which predicts tweet popularity using network information. Models based on the theory of point processes [Zhao et al. (2015), Kobayashi and Lambiotte (2016)], which do not require network information, were also shown to have good prediction performances. The model proposed by Mishra, Rizoiu and Xie (2016), on the other hand, combines the point process models with feature based approaches to predict the popularity. Other models like the growth-adoption model by Lymperopoulos (2016), the spatial-temporal heterogeneous Bass model by Yan et al. (2016), and the concept drift model by Li et al. (2016), were all proposed for the purpose of popularity predictions.

Some models employed on other online social networks (OSNs) with the same purpose of popularity predictions are also relevant as the proposed methodologies may be applicable to predictions on Twitter network. Notably, Agarwal, Chen and Elango (2009) proposed a dynamic linear regression model to predict the click-through rate for Today Module on Yahoo! Front Page. Activities on platforms like Youtube and Digg were also modelled, for instance using linear regression models [Szabo and Huberman (2010)] and classification models [Ahmed et al. (2013)]. Other related works include the reinforced Poisson model [Gao, Ma and Chen (2015)] applied on Sina Weibo and the model by Wu et al. (2016) that incorporates temporality and seasonality, applied on Flickr image data.

Recently Zhao et al. (2015) proposed a model termed the SEISMIC (Self-Exciting Model of Information Cascades) to model the retweeting dynamics on Twitter. The model describes the retweet intensity of a tweet, or the expected number of retweets per unit time, as a product of the infectivity of the original tweet and the accumulated excitation effect of all previous retweets. Zhao et al. (2015) estimated the infectivity as a function of time using a kernel smoothing estimator, and the excitation function, or *memory kernel*, using a graphical approach under the assumption that some retweeting processes follow an inhomogeneous Poisson process with the excitation function as its intensity function. They also proposed to predict the future popularity of a tweet based on calculating the expected number of future retweets by first assuming that the infectivity remains constant since the censoring time, and a subsequent ad hoc adjustment to the expectation to incorporate the decaying trend of the infectivity. They reported that the predictions of

tweet popularity using their approach outperformed those based on competing approaches, under several performance measures. Kobayashi and Lambiotte (2016) proposed another model termed the TiDeH (Time-Dependent Hawkes) model, which models the retweet intensity similar to the SEISMIC, and estimated the infectivity and memory kernel using similar nonparametric kernel smoothing estimators. Kobayashi and Lambiotte (2016) fitted a circadian rhythmic function to the nonparametrically estimated infectivity function up to the censoring time, and extrapolated it beyond the censoring time to predict the number of future retweets. With certain choices of the smoothing parameters, the tweet popularity predictions based on the TiDeH model are superior to those based on the SEISMIC, especially on long cascades. However, Kobayashi and Lambiotte's approach requires sufficiently long observation time on a retweet sequence to have reliable estimation of the infectivity function, and the prediction performance depends critically on the window size parameter used in the estimation of the infectivity function.

In this work, we propose a marked self-exciting point process model to capture the retweeting dynamics and to predict tweet popularity. Our model is motivated by the SEISMIC and the TiDeH model, and bears some similarities to them. However, our model has some important advantages. First, the intensity process in our model has a linear form similar to that of the original self-exciting process of Hawkes (1971), and therefore the resulting point process is interpretable as a cluster Poisson process, which means our model can be simulated using a cascading algorithm similar to that used for the efficient simulation of Hawkes processes. Second, the estimation of our model and the assessment of the goodness-of-fit can be implemented using principled approaches from point process theory, and the predictions based on our model can also be done properly by exploiting the probabilistic properties of the model, without resorting to ad hoc assumptions such as those needed by the SEISMIC. Moreover, our model is found to be able to capture the retweeting dynamics and make accurate popularity predictions based on much shorter observation times than those required by the TiDeH model.

The rest of the paper is organized as follows. In Section 2 we describe the tweet data and previous models for tweet popularity predictions which motivated our work, and show how our model is built, how we estimate the parameters and evaluate the goodness-of-fit, and finally how to make predictions. In Section 3 we apply the proposed model to the tweet data and compare the prediction performance of our model with that of the SEISMIC and the TiDeH model, to show the superior performance of our model. Finally, in Section 4, we conclude the paper with a discussion.

2. The data, the model, and the methodology.

2.1. *Twitter data*.   The data[1] which motivated our work and which will be used to demonstrate our modelling and prediction methodology is that used recently by

---

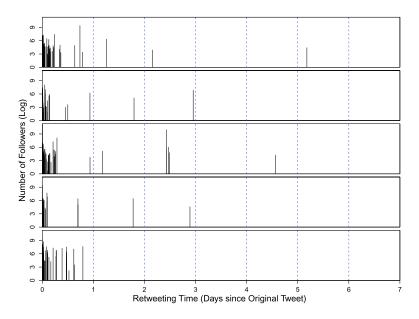[1]Available from http://snap.stanford.edu/seismic/.

FIG. 1. *Times of retweets and the corresponding numbers of followers of the retweeting accounts on the log scale, for five randomly selected retweet cascades from the training data set.*

Zhao et al. (2015). The data contains a total of 166,069 reasonably popular tweets published from October 7 to November 7, 2011, each with at least 49 retweets within seven days of publishing. For each tweet, the data includes the Twitter ID of the original tweeter, the posting times of the original tweet and all the retweets within seven days, and the numbers of followers of the original poster and of the retweeters. Following Zhao et al. (2015), we use data on the 71,815 tweets published in the first seven days of the study period as training data and the remaining 94,254 tweets published in the next eight days as test data. See Figure 1 for 5 randomly selected retweet cascades from the training data set.

Note that the data lacks the complete Twitter network information, that is, for a retweet, the data only has its publishing time and the number of followers of the retweeting account, without information on whether the original tweet or any previous retweet is being retweeted. This implies that the methodology of Zaman, Fox and Bradlow (2014), which assumes the complete Twitter network information, does not apply here. The contents of the original tweets and of the retweets are also not included in the data, and therefore methodologies depending on features of the tweet contents or features of the posters (other than their numbers of followers) do not apply either.

From Figure 1 we observe that the retweets tend to occur in clusters or bursts. This suggests that self-exciting processes like the SEISMIC of Zhao et al. (2015) are potentially suitable for such data. In the next subsection we present our model

for the retweeting dynamics, which is a marked self-exciting process model similar to that of the SEISMIC.

2.2. *Model formulation.* Let $(\tau_i, m_i), i = 0, 1, \ldots$ be a marked point process where $0 = \tau_0 < \tau_1 < \cdots$ denote the *event times* and $m_0, m_1, \ldots$ denote the respective *event marks*. In the context of information diffusion modelling on Twitter, the event times shall refer to the retweeting times, except that $\tau_0 = 0$ denotes the posting time of the original tweet, and the event marks refer to the numbers of followers of the retweeting (or tweeting, in the case of m_0) accounts. Let $N(t) = \sum_{i=1}^{\infty} \mathbb{I}\{\tau_i \leq t\}, t \geq 0$ be the associated counting process of retweets, and $\mathcal{F} = \{\mathcal{F}_t; t \geq 0\}$, with $\mathcal{F}_t = \sigma\{N(t), m_0, (\tau_j, m_j), j = 1, \ldots, N(t)\}$, be the natural filtration of the marked point process. The *(conditional) intensity process* of N relative to the filtration $\mathcal{F}$ is an $\mathcal{F}$-predictable process $\lambda(t), t \geq 0$, such that $M(t) = N(t) - \int_0^t \lambda(s)\, ds, t \geq 0$ is a mean zero $\mathcal{F}$-martingale. In an informal but intuitive notation, the intensity can be written as $\lambda(t) = \mathbb{E}[dN(t)|\mathcal{F}_{t-}]/dt$, from which we note that the intensity at any time point is the expected number of events per unit time given the history of the process prior to that time point.

As the evolution of a point process over time is fully determined by its intensity process, a commonly used approach to specify a point process model is to specify the form of the dependence of its intensity process on the prior-t history of the process $\mathcal{F}_{t-}$. In particular, the SEISMIC of Zhao et al. (2015) assumes that the intensity of the retweeting process $N(t)$ takes this form,

$$
(1) \qquad \lambda(t) = p(t) \sum_{i=0}^{N(t-)} m_i \phi(t - \tau_i), \qquad t > 0,
$$

where $p(t)$ is an unspecified positive function, called the infectivity function, which typically decreases in t, and $\phi(\cdot)$ is a positive function called the memory kernel. Zhao et al. (2015) proposed to estimate the infectivity function $p(\cdot)$ nonparametrically using a kernel smoothing estimator with a triangular kernel. To estimate the memory kernel, they assumed that it is of a power law decaying form and that 15 "carefully chosen" retweet cascades follow inhomogeneous Poisson processes with intensity functions proportional to the memory kernel. They then estimated the parameters using histogram and complementary cumulative distribution function plots of the retweeting times in those 15 cascades. As for the TiDeH model of Kobayashi and Lambiotte (2016), it assumed an intensity process of the same form as (1), except the further assumption that the infectivity function $p(\cdot)$ is also parametric, and takes a dampened circadian oscillation form. To estimate the infectivity function $p(\cdot)$, Kobayashi and Lambiotte (2016) proposed a two-step approach where a preliminary estimate $\hat{p}(\cdot)$ was obtained first using a kernel method and then the parametric form of $p(\cdot)$ was fitted to the preliminary estimate by a least squares method.

The point process model we propose in the current work for the purpose of retweeting dynamics modelling is given by

$$(2) \qquad \lambda(t) = \nu(t) + \sum_{i=1}^{N(t-)} \omega(\tau_i, m_i, t - \tau_i),$$

where $\nu(\cdot)$ is the *baseline intensity function*, with $\nu(t)$ denoting the part of the event intensity at time t that is due to the initial event at time 0; and $\omega(\cdot, \cdot, \cdot)$ is the *excitation function*, with $\omega(\tau, m, t - \tau)$ denoting the impact of an event at time τ with mark m on the event intensity at time t, where t is the time since the initial tweet was posted. Furthermore, both the baseline intensity and the excitation functions are time-dependent and take multiplicatively separable forms as follows:

$$\nu(t) = \alpha \phi(t),$$
(3)
$$\omega(\tau, m, t - \tau) = p(\tau) r(m) \phi(t - \tau).$$

Here $\alpha > 0$ is a constant giving the direct excitation effect of the original tweet, that is, how many retweets it is expected to generate directly. The function $\phi(\cdot)$ is called the *memory kernel function*, which describes how the excitation effect due to the original tweet or a retweet is distributed over time. Similar to Zhao et al. (2015), we require $\phi(\cdot)$ to be a probability density function, so that $\phi(\cdot) \geq 0$ and $\int_0^\infty \phi(t) \, dt = 1$. The function $p(\cdot)$ indicates how the "infectivity" of a retweet varies over time and is also called the *infectivity function*, although its influence on the intensity process $\lambda(t)$ is different than that of the infectivity function $p(\cdot)$ in (1). For identifiability, we assume that $p(0) = 1$. The function $r(\cdot)$ is called the *impact function*, and describes the total excitation effect of a retweet attributed to the number of followers of the retweeter. Note, we do not require $\alpha = r(m_0)$, to allow for the potentially different influences of the original tweet and of the retweets.

More specifically, the functions in (3) are assumed to take the following parametric forms:

$$p(\tau; \beta) = e^{-\beta \tau},$$
(4)
$$r(m; \gamma) = \gamma \log(m + 1),$$
$$\phi(t; \delta) = \frac{\delta_2 (\delta_1 - 1)}{\delta_1} \left(1 + \frac{\delta_2 t}{\delta_1} \right)^{-\delta_1}$$

for parameters $\beta \geq 0$, $\gamma \geq 0$, $\delta_1 > 1$, and $\delta_2 > 0$. Here we have adopted an exponential decay form for the infectivity function, based on the intuition that the infectivity, or newsworthiness of a retweet should decay very quickly over time. We further assume that the impact function is linear in the number of followers on a log scale, rather than on the original scale as in Zhao et al. (2015), because of the high degree of right skewness for the distribution of the number of followers [Cha et al. (2010), Kwak et al. (2010), Bakshy et al. (2011)]. Our choice of the

power law decay form for the memory kernel is motivated by Zhao et al. (2015) and the empirical findings of the heavy tailed distributions for the human response time in social networks, reported in the literature [Barabasi (2005), Crane and Sornette (2008), Zaman, Fox and Bradlow (2014)]. Finally, as in Zhao et al. (2015) and Kobayashi and Lambiotte (2016), we also assume the event marks m_i are i.i.d. with a common density function $f(\cdot)$ relative to a suitable reference measure on the space $\mathcal{M}$ of event marks, and moreover, m_i is independent of τ_i and $\mathcal{F}_{\tau_i-}$ for all i. As the excitation function associated with an event is allowed to depend on the time of that event, the model will be called a *Marked Self-Exciting Process with Time-Dependent Excitation Function*, or MaSEPTiDE for short.

At this point we emphasize an important difference between the MaSEPTiDE we propose in this work and the SEISMIC of Zhao et al. (2015). From (2), we note that, unlike the SEISMIC, the MaSEPTiDE has an intensity process that is of a linear form similar to the self-exciting process of Hawkes (1971), whose intensity process takes the form

$$\lambda(t) = \nu + \sum_{i=1}^{N(t-)} g(t - \tau_i).$$

In fact, if we choose $p(\tau) \equiv 1$ and $r(m) \equiv r$ for a constant r in (3), then (2) reduces to the time-varying version of the Hawkes process considered by Chen and Hall (2013, 2016). The linear structure of the intensity process implies that the MaSEPTiDE can also be interpreted as a Poisson cluster process, as the original Hawkes process or the generalized version with a time-varying background intensity. By this interpretation, immigrants arrive according to a marked inhomogeneous Poisson process with its intensity function equal to the baseline intensity function $\nu(\cdot)$, and event mark distributed according to the density function $f(\cdot)$. Once an immigrant with mark m arrives at τ, it starts to independently produce children according to a marked inhomogeneous Poisson process with intensity function $\omega(\tau, m, \cdot) = p(\tau)r(m)\phi(\cdot)$ and event marks distributed according to $f(\cdot)$, so that the total number of children is Poisson distributed with mean $\int_0^\infty \omega(\tau, m, t)\, dt = p(\tau)r(m)$, and given the total number of children, the waiting times to births of the children are i.i.d. with common density function $\phi(\cdot)$ if the order of births is ignored. Moreover, once an offspring of any generation is born, say at time τ' and with mark m', it starts to independently produce children of its own according to a similar marked inhomogeneous Poisson process with intensity function $\omega(\tau', m', \cdot)$ and event marks distributed according to $f(\cdot)$. The events of the MaSEPTiDE process by time t consist of all immigrants and offspring of any generation that have arrived by time t. This Poisson cluster process interpretation implies an efficient recursive cascading algorithm to simulate the MaSEPTiDE process as described in Section 2.5, which has important implications for simulation based predictions by the MaSEPTiDE process.

Because of the Poisson cluster interpretation, the memory kernel function $\phi(\cdot)$ in the MaSEPTiDE can also be called the *offspring density function*, and the function $p(\cdot)r(\cdot)$ might be interpreted as the *branching ratio function* which specifies how the branching ratio, that is, the average number of direct (or generation 1) offspring from an individual (be it an immigrant or an offspring), depends on the birth time and event mark of the individual. In contrast, the functions $p(\cdot)$ and $\phi(\cdot)$ in the SEISMIC or the TiDeH model do not permit such a neat interpretation.

It might also be of interest to note the difference between the treatments of the background intensity in the MaSEPTiDE model and in the Hawkes process model with a general time-varying background intensity. In the former model, we require the baseline intensity function to be proportional to the memory kernel $\phi(\cdot)$, while in the latter, the background intensity and the memory kernel can take different shapes. The advantage of our treatment is that it leads to a more parsimonious model, while the time-varying background intensity model can easily accommodate nonstationarity, such as that due to the diurnal patterns of human activity levels.

2.3. *Parameter estimation.*   Before we can use the MaSEPTiDE model for future events prediction, we need to first estimate the model parameters. Since the event marks are assumed to be i.i.d., their distribution can simply be estimated by the empirical distribution of m_i, for $i = 1, \ldots, N(T)$. The main estimation problem is to estimate the parameter vector $\theta = (\alpha, \beta, \gamma, \delta_1, \delta_2)^\top$. To this end, we shall use the maximum likelihood (ML) approach. By the point process theory [Daley and Vere-Jones (2003), Proposition 7.3.III] the likelihood of the MaSEPTiDE process based on observations over the interval $[0, T]$, where T denotes the censoring time, takes the following form:

$$(5) \qquad L(\theta) = \left\{ \prod_{i=1}^{N(T)} \lambda(\tau_i) \right\} \exp\left(- \int_0^T \lambda(t)\, dt \right) \prod_{i=1}^{N(T)} f(m_i),$$

where $\lambda(\cdot)$ depends on the parameters through (2)–(4), and $f(\cdot)$ denotes the event mark density, which is assumed to be free of the parameters θ.

To compute the ML estimator of the parameter vector θ using general-purpose numerical optimization routines, the efficient evaluation of the likelihood function or its logarithm is very important. For this purpose, we need to be able to evaluate the definite integral of the intensity function in (5) efficiently. Due to the linear structure of the intensity function, the integral of the intensity function can be shown to take an explicit form similar to the intensity function itself, and therefore can be exactly computed without resorting to numerical quadrature routines. To show this, it is convenient to use the random measure interpretation of a marked point process. That is, we interpret

$$N(d\tau, dm) = \sum_{i=1}^{\infty} \delta_{(\tau_i, m_i)}(d\tau, dm)$$

as a random measure on $[0, \infty) \times \mathcal{M}$, so that the intensity in (2) can be written as

$$\lambda(t) = \nu(t) + \sum_{i=1}^{N(t-)} \omega(\tau_i, m_i, t - \tau_i) = \nu(t) + \int_{(0,t) \times \mathcal{M}} \omega(\tau, m, t - \tau) N(d\tau, dm).$$

Therefore, by Fubini's theorem, a change of variables, and the assumed forms of the functions ν, ω and ϕ, we have

$$\begin{aligned}
\int_0^T \lambda(t)\, dt &= \int_0^T \nu(t)\, dt + \int_0^T \int_{(0,t) \times \mathcal{M}} \omega(s, m, t - s) N(ds, dm)\, dt \\
&= \int_0^T \nu(t)\, dt + \int_{(0,T) \times \mathcal{M}} \int_s^T \omega(s, m, t - s)\, dt\, N(ds, dm) \\
&= \int_0^T \nu(t)\, dt + \int_{(0,T) \times \mathcal{M}} \int_0^{T-s} \omega(s, m, t)\, dt\, N(ds, dm) \\
&= \alpha \Phi(T) + \sum_{i=1}^{N(T-)} p(\tau_i) r(m_i) \Phi(T - \tau_i),
\end{aligned}$$

(6)

where $\Phi(t) = \Phi(t; \delta) = \int_0^t \phi(s; \delta)\, ds = 1 - (1 + \delta_2 t / \delta_1)^{-\delta_1 + 1}$, $t \geq 0$.

From the separable form of the likelihood function in (5) and the assumption that the event mark distribution does not depend on the parameter vector θ, the ML estimation of the parameters θ can be based on maximizing the logarithm of the part of the likelihood that does not involve $f(\cdot)$, that is

$$\ell(\theta) = \sum_{i=1}^{N(T)} \log \lambda(\tau_i) - \int_0^T \lambda(t)\, dt.$$

In practice, the maximization can be done numerically using various general-purpose optimization routines. In our numerical experiments, we have used the downhill simplex method of Nelder and Mead (1965), which is the default method used by the function `optim` in the R software environment for statistical computing [R Core Team (2016)].

2.4. *Goodness-of-fit assessment.* The assessment of the goodness-of-fit of models to historical data can guide us to seek models that can describe the observed data well and therefore serves as the basis of predictions for future observations. To assess the goodness-of-fit of the MaSEPTiDE, we shall use the residual point process approach based on Papangelou's random time change theorem [Daley and Vere-Jones (2003), Theorem 7.4.I]. By the time change theorem, with $\Lambda(t) = \int_0^t \lambda(s)\, ds$ denoting the cumulative intensity process, the transformed process $N(\Lambda^{-1}(t))$ is a Poisson process with unit rate or equivalently, the random times $\Lambda(\tau_i)$, $i = 1, 2, \ldots$, will be the event times of a unit rate Poisson process. Therefore, if the MaSEPTiDE with the parameters θ set to their ML es-

timates $\hat{\theta}$ is a sufficient model for the observed event times up to the censoring time T, then the transformed event times, $\hat{\Lambda}(\tau_i)$, $i = 1, \ldots, N(T)$ should be approximately equal in distribution to the event times of a unit rate Poisson process up to time $\hat{\Lambda}(T)$, where $\hat{\Lambda}(t)$, $t > 0$ is the plugin estimate of the cumulative intensity $\Lambda(t; \theta) = \int_0^t \lambda(s; \theta)\, ds$, that is,

$$\hat{\Lambda}(t) = \Lambda(t; \hat{\theta}) = \hat{\alpha}\Phi(t; \hat{\delta}) + \sum_{i=1}^{N(t-)} p(\tau_i; \hat{\beta})r(m_i; \hat{\gamma})\Phi(t - \tau_i; \hat{\delta}),$$

with $p(\cdot)$ and $r(\cdot)$ defined in (4), and $\Phi(\cdot)$ defined as in (6). As the conditional distribution of the event times of a Poisson process in a fixed interval given the total number of events in the interval is equal in distribution to the order statistics of the same number of i.i.d. random variables uniformly distributed in the interval, to assess the goodness-of-fit of the MaSEPTiDE (or any point process model specified via the intensity process), we can assess the uniformity of the transformed event times $\hat{\Lambda}(\tau_i)$, $i = 1, \ldots, N(T)$, in the interval $(0, \hat{\Lambda}(T)]$, informally using graphical approaches such as the histogram or the Q-Q (quantile–quantile) plots, or formally using statistical tests like the K-S (Kolmogorov–Smirnov) test. A similar residual analysis was performed by Ogata (1988) to assess the goodness-of-fit of point process models on earthquake data.

2.5. *Predicting future number of events.* Given observations up to T, to predict the number of events from T to a future time point $\tilde{T} > T$, one commonly uses its conditional expectation or its conditional median, which is optimal relative to the mean squared error or the mean absolute error accordingly [Gneiting (2011)]. To obtain the conditional expectation, we can use either a solve-the-equation approach or a simulation based approach. The former approach involves deriving a functional equation satisfied by the conditional expectation as a function of a future time point, solving the equation, and evaluating the solution function at the desired time point. The latter approach involves simulating the sample path of the MaSEP-TiDE on the time interval $(T, \tilde{T}]$ conditional on the observations up to time T for a large number of times, counting the number of events on each simulated sample path, and using the average of the simulated event counts to approximate its expectation. While the first approach is computationally less expensive, the solution of the functional equation is not always easy to obtain. For the second approach, although it is relatively less efficient, especially if the process to be simulated has a large expected number of events, it is more robust than the first approach. To obtain the conditional median, the only option seems to be a simulation based approach, which involves simulating the conditional sample path of the MaSEPTiDE process a large number of times and extracting the median of the resultant empirical distribution of the number of events in the prediction interval.

Both solve-the-equation approach and simulation based approach rely on the observation that, conditional on the history of the MaSEPTiDE process up to time

T, its future evolution is the same as that of another MaSEPTiDE process with a different baseline intensity function and a similar excitation function. Let $\tilde{N}(t) = N(T + t) - N(T)$, for $t \geq 0$, $\tilde{\tau}_j = \tau_{N(T)+j} - T$, $\tilde{m}_j = m_{N(T)+j}$ for $j = 1, 2, \ldots$, and $\tilde{\mathcal{F}}_t = \mathcal{F}_{T+t}$, $t \geq 0$. Then, the $\tilde{\mathcal{F}}$-intensity process of $\tilde{N}(t)$ is given by

$$\tilde{\lambda}(t) = \lambda(T + t) = \nu(T + t) + \sum_{j=1}^{N(T)} \omega(\tau_j, m_j, T + t - \tau_j)$$

$$+ \sum_{j=N(T)+1}^{N(T+t-)} \omega(\tau_j, m_j, T + t - \tau_j)$$

$$= \tilde{\nu}(t) + \sum_{j=1}^{\tilde{N}(t-)} \tilde{\omega}(\tilde{\tau}_j, \tilde{m}_j, t - \tilde{\tau}_j),$$

where $\tilde{\nu}(\cdot)$ denotes the function

$$(7) \qquad \tilde{\nu}(t) = \nu(T + t) + \sum_{j=1}^{N(T)} \omega(\tau_j, m_j, T + t - \tau_j),$$

and $\tilde{\omega}(\cdot, \cdot, \cdot)$ denotes the function

$$(8) \qquad \tilde{\omega}(\tau, m, t) = \omega(T + \tau, m, t) = p(T + \tau)r(m)\phi(t) \equiv \tilde{p}(\tau)r(m)\phi(t).$$

Therefore, $\tilde{N}(t)$, $t \geq 0$ is a MaSEPTiDE process with baseline intensity function $\tilde{\nu}$ and excitation function $\tilde{\omega}$ given as above in (7) and (8) respectively. The excitation function $\tilde{\omega}$ has a similar separable form as ω, with r and ϕ the same as before, and infectivity function $\tilde{p}$ equal to a time shift of the previous infectivity function, that is, $\tilde{p}(\tau) = p(T + \tau)$.

To calculate the expected number of events $\mathbb{E}[N(\tilde{T}) - N(T)|\mathcal{F}_T]$ without resorting to simulations, we first note from the definition of the conditional intensity that

$$(9) \qquad \begin{aligned} \mathbb{E}[N(\tilde{T}) - N(T)|\mathcal{F}_T] &= \mathbb{E}[\tilde{N}(\tilde{T} - T)|\mathcal{F}_T] = \mathbb{E}\left[\int_0^{\tilde{T}-T} \tilde{\lambda}(s)\,\mathrm{d}s \bigg| \mathcal{F}_T\right] \\ &= \int_0^{\tilde{T}-T} \mathbb{E}[\tilde{\lambda}(s)|\mathcal{F}_T]\,\mathrm{d}s = \int_0^{\tilde{T}-T} \bar{\lambda}(s)\,\mathrm{d}s, \end{aligned}$$

with $\bar{\lambda}(s) = \mathbb{E}[\tilde{\lambda}(s)|\mathcal{F}_T]$ denoting the mean intensity function of $\tilde{N}(t)$ given $\mathcal{F}_T$. By the independence between event marks and previous event times, we have

$$\begin{aligned} \bar{\lambda}(t) &= \mathbb{E}[\tilde{\lambda}(t)|\mathcal{F}_T] \\ &= \mathbb{E}\left[\tilde{\nu}(t) + \int_{(0,t)\times\mathcal{M}} \tilde{\omega}(\tau, m, t - \tau)\tilde{N}(\mathrm{d}\tau, \mathrm{d}m) \bigg| \mathcal{F}_T\right] \end{aligned}$$

(10)

$$= \mathbb{E}\left[\tilde{\nu}(t) + \int_{(0,t)\times\mathcal{M}} \tilde{p}(\tau)r(m)\phi(t-\tau)\tilde{\lambda}(\tau)\,\mathrm{d}\tau\,\mathrm{d}F(m)\Big|\mathcal{F}_T\right]$$

$$= \tilde{\nu}(t) + \int_{\mathcal{M}} r(m)\,\mathrm{d}F(m)\int_0^t \tilde{p}(\tau)\phi(t-\tau)\mathbb{E}[\tilde{\lambda}(\tau)|\mathcal{F}_T]\,\mathrm{d}\tau$$

$$= \tilde{\nu}(t) + R\int_0^t \tilde{p}(\tau)\phi(t-\tau)\bar{\lambda}(\tau)\,\mathrm{d}\tau,$$

where we have also used $\tilde{N}(\mathrm{d}\tau,\mathrm{d}m)$ to denote the associated random measure again, and F denotes the distribution of the i.i.d. event marks, while $R = \mathbb{E}[r(m_i)] = \int_{\mathcal{M}} r(m)\,\mathrm{d}F(m)$ is the expected total excitation effect due to an event.

In general, we need to solve the integral equation in (10) numerically to obtain $\bar{\lambda}(t)$ on $[0, \tilde{T} - T]$ and use it in finding the conditional expectation of the number of events in (9). One method to solve (10) is to approximate $\bar{\lambda}(t)$ by a flexible parametric function and identify the parameters by requiring both sides of the equation to be equal or approximately equal at sufficiently many points in the interval $[0, \tilde{T} - T]$. Examples of the flexible parametric functions to approximate $\bar{\lambda}(t)$ include a B-spline function with a specified order and knot sequence, or a truncated Fourier series. In both cases, the unknown parameters of the approximating function can be obtained by solving a linear equation of the unknown parameters. In practice, we would try approximating functions with increasing flexibility until convergence in the solution is achieved. We select the B-spline function in this work as a method to find $\bar{\lambda}(t)$ for its ease of implementation and computational stability. Specifically, we let $B(t) = (B_1(t), \ldots, B_k(t))^\top$ denote the set of B-spline basis functions of a certain order on the interval $(0, \tilde{T} - T)$, and assume that $\bar{\lambda}(t) \approx B(t)^\top \eta$ for a k-vector η. Plugging this into (10), we have the following equation of η:

(11)

$$B(t)^\top \eta = \tilde{\nu}(t) + \left\{ R\int_0^t \tilde{p}(\tau)\phi(t-\tau)B(\tau)^\top \eta\,\mathrm{d}\tau \right\}.$$

To solve (11) for η, we evaluate both sides of (11) at sufficiently many ($\geq k$) t values in the interval $(0, \tilde{T} - T)$, and solve the resulting overdetermined linear system using the method of least squares to get η. Once η is obtained, the predicted value is calculated as

$$\{N(\tilde{T}) - N(T)\}_{\mathrm{pred}} = \left(\int_0^{\tilde{T}-T} B(t)\,\mathrm{d}t \right)^\top \eta.$$

In evaluating the integrals in (11), we often need to use numerical quadrature routines. In our case, we have used the R function `integrate` for this purpose.

To simulate the MaSEPTiDE process $\tilde{N}$ over the interval $(0, \tilde{T} - T)$, we can use the following cascading algorithm, which is a generalization of that used for the simulation of nonstationary self-exciting point processes [Chen and Hall (2013, 2016)]. A similar algorithm has been used by Chen and Stindl (2018) to simulate renewal Hawkes processes.

1. Simulate an inhomogeneous Poisson process N^0 with time-varying intensity function $\tilde{\nu}(t)$ on $(0, \tilde{T} - T]$ and denote the event times of N^0 by τ_j^0, $j = 1, \ldots, N^0(\tilde{T} - T)$.

2. Generate the associated event marks m_j^0 independently from the event mark distribution F and call the events (τ_j^0, m_j^0), $j = 1, \ldots, N^0(\tilde{T} - T)$ generation 0 events.

3. For each generation 0 event (τ_j^0, m_j^0), simulate an inhomogeneous marked Poisson process N_j^1, with intensity function $\tilde{\omega}(\tau_j^0, m_j^0, \cdot)$ and event mark distribution F, on the interval $(0, \tilde{T} - T - \tau_j^0]$ and denote the corresponding events by (τ_{jk}^1, m_{jk}^1), $k = 1, \ldots, N_j^1(\tilde{T} - T - \tau_j^0)$. The collection of events $\{(\tau_j^0 + \tau_{jk}^1, m_{jk}^1);$ $k = 1, \ldots, N_j^1(\tilde{T} - T - \tau_j^0), j = 1, \ldots, N^0(\tilde{T} - T)\}$ are referred to as generation 1 events.

4. Continue generating events of generations 2, 3, $\ldots$ similarly on intervals of decreasing lengths, until a generation has no events.

5. The events of all generations are pooled together to form the collection of all events of the MaSEPTiDE $\tilde{N}$ process on the interval $(0, \tilde{T} - T]$.

The algorithm shown above requires the simulation of inhomogeneous Poisson processes, which can be achieved using the thinning algorithm of Lewis and Shedler (1979). In our numerical experiments, we have used the R implementation simPois from the IHSEP package. Our implementation of the above cascading algorithm is based on a simple modification of the function simHawkes1 from the R package IHSEP.

To predict the number of events in the interval $(T, \tilde{T}]$, we simulate the sample path of the process $\tilde{N}(t)$ over the interval $(0, \tilde{T} - T]$ for a large number (say 100) of times, and count the number of events on each simulated sample path. The mean or median of these simulated event numbers will then be our point prediction of the number of events of the MaSEPTiDE process N in the interval $(T, \tilde{T}]$.

In practice, when we use the fitted model to make predictions, whether by using the solve-the-equation approach or by using the simulation based approach, the unknown functions $\tilde{\nu}$ and $\tilde{\omega}$, and the event mark distribution F need to be replaced by their respective estimators. In our numerical experiments, we use the plugin estimators $\tilde{\nu}(t; \hat{\theta})$ and $\tilde{\omega}(\cdot, \cdot, \cdot; \hat{\theta})$ for $\tilde{\nu}$ and $\tilde{\omega}$, and the empirical distribution function $\hat{F}$ of the event marks $m_1, \ldots, m_{N(T)}$ for F. One implication is that the constant R in (10) is set to $\hat{R} = \int_{\mathcal{M}} r(m) \, d\hat{F}(m) = \sum_{i=1}^{N(T)} r(m_i)/N(T)$. Finally, we note that, if the target of prediction is the total number of events of the process N in the interval $(0, \tilde{T}]$, then we simply add the observed number in the interval $(0, T]$, that is, $N(T)$, to the predicted number in the interval $(T, \tilde{T}]$.

3. Application to the tweet data. In this section, we report the results of applying the proposed model and inference methodologies to the tweet data. The

performance of our prediction methods is also compared to those of the SEISMIC and the TiDeH model.

3.1. *The model fit.* We fitted the MaSEPTiDE model to the 71,815 retweet cascades in the training data set described in Section 2.1 with different censoring times, using the maximum likelihood method described in Section 2.3. The estimated parameter values with censoring time of seven days are highly skewed, with the median estimates of α, β, γ, δ_1, and δ_2 equal to 48.349, 0.072, 7.209, 1.416, and 0.007 respectively. To have some idea about the typical parameter values found in practice, we display in Table 1 the estimated parameter values for the five sample cascades shown in Figure 1, together with their final popularity.

The estimated values of the parameter β suggest very fast decays of infectivity, with the times taken for the infectivity to drop to 1% of the initial levels vary from about 19 seconds [$\log(100)/0.246 = 18.7$ seconds] in sample cascade 3 to about 4 minutes [$\log(100)/0.019 = 242.4$ seconds] in sample cascade 5. While the estimated values of the shape parameter of the memory kernel δ_1 are more or less similar to each other, the scale parameter δ_2 is substantially more variable. In particular, the extremely small $\hat{\delta}_2$ value of 0.001 in sample cascade 3 implies a very long range memory effect, which, together with a relatively large $\hat{\beta}$ value, suggest that the later retweets are more likely to be generated by the original tweet or retweets within the first few seconds of the original tweet (if any), while in sample cascade 1, where the $\hat{\delta}_2$ value is 0.173, the later retweets are more likely to be generated by more recent retweets. The estimated values of the scale parameter α of the baseline intensity function, together with the values of the δ parameters and the final popularity, suggest highly variable proportions of generation 0 retweets, ranging from 3.4% [$= 5.711\Phi(\tilde{T}; 1.254, 0.173)/159$] in sample cascade 1 to nearly 100% [$= 58.136\Phi(\tilde{T}; 1.490, 0.001)/55$] in sample cascade 3. Also, the estimated γ values on the five sample cascades have quite substantial variation, with the increase in the excitation effect associated with one unit increase in the number of followers (on the log scale) of a retweeting account vary from 1.144 to 6.351 units.

TABLE 1
Fitted parameter values on the sample cascades shown in Figure 1

Sample cascade	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$	$\hat{\delta}_1$	$\hat{\delta}_2$	$N(\tilde{T})$
1	5.711	0.024	1.455	1.254	0.173	159
2	3.075	0.021	6.351	1.414	0.029	85
3	58.136	0.246	1.144	1.490	0.001	55
4	8.209	0.031	2.095	1.444	0.040	74
5	4.173	0.019	5.049	1.229	0.046	89

TABLE 2
The percentages of cascades in the training data set where the MaSEPTiDE model passes the goodness-of-fit test at different significance levels and censoring times

Significance level	Censoring time (hours)						
	2	**4**	**6**	**8**	**10**	**12**	**168**
0.01	92.0%	88.2%	85.8%	84.2%	82.8%	81.8%	74.9%
0.05	89.3%	84.7%	81.9%	80.1%	78.5%	77.5%	69.2%

By the goodness-of-fit assessment method described in Section 2.4, we tested the uniformity of the point process residuals $\hat{\Lambda}(\tau_i)$ over the interval $(0, \hat{\Lambda}(T)]$ using the K-S test. At significance levels of 0.01 and 0.05 with different censoring times, the percentages of the 71,815 cascades where the estimated MaSEPTiDE model passes the residual uniformity test are shown in Table 2.

From this table we note that the percentage of cascades from which the estimated model passes the test decreases when the censoring time increases. This is to be expected as the amount of data increases with the censoring time, implying that the difficulty of finding a fitting model also increases. At significance level of 0.01, when fitted to the complete cascade data, that is, with the censoring time of 168 hours (seven days), the MaSEPTiDE model passes the goodness-of-fit test on roughly 75% of the cascades. By the censoring time of 12 hours, the MaSEPTiDE model passes the goodness-of-fit test on the majority (approximately 82%) of the cascades. Given that the majority of the retweets, or 80% on average, have already happened within 12 hours of the posting of the original tweets, we conclude that the MaSEPTiDE model is able to describe the retweeting dynamics reasonably well.

3.2. *Popularity prediction.* For each of the 94,254 tweets in the test data set, we applied the fitted MaSEPTiDE model with the retweet cascades censored at different times to predict their final popularity, using the prediction methods discussed in Section 2.5. For the purpose of comparison, we also obtained the predictions based on the SEISMIC of Zhao et al. (2015) and the TiDeH model of Kobayashi and Lambiotte (2016). We only report the results of comparison with these two methods, because they were found to outperform other methods in the literature, such as those reported in Crane and Sornette (2008), Agarwal, Chen and Elango (2009), Szabo and Huberman (2010), and Gao, Ma and Chen (2015), both in our own numerical experiments and in the works of Zhao et al. (2015) and Kobayashi and Lambiotte (2016).

Our point prediction of the "final" popularity of a tweet, or the total number of retweets by time $\tilde{T} = 7$ days, using the MaSEPTiDE model estimated with the retweet cascade observed up to the censoring time T, is given by $N(\tilde{T})_{\text{pred}} =$

$N(T) + (N(\tilde{T}) - N(T))_{\text{pred}}$, where $(N(\tilde{T}) - N(T))_{\text{pred}}$ is obtained either as the conditional expectation using the solve-the-equation approach or the simulation based approach, or as the conditional median using the simulation based approach. Our numerical experiments have confirmed that in the case of conditional expectation, the two approaches produce identical predictions up to a negligible numerical error, as expected. When using the simulation based approach, we note that, for some very popular tweets, the retweeting cascades are very long and the numbers of retweeting events to be simulated in the prediction intervals are very large, and therefore simulations can take a long time to complete. This issue was also noted by Kobayashi and Lambiotte (2016). A trick we used to mitigate this issue is to simulate the process $\tilde{N}$ with a smaller baseline intensity function, say $\tilde{\nu}(\cdot)/S$ with $S = 100$ or larger, and inflate the simulated event numbers by the factor S. For the majority of the cascades, a moderately large number of simulation replications, such as 100 or even 50, was enough to produce a prediction consistent with that by the solve-the-equation approach. The same set of simulations were used to calculate the median based prediction.

To assess the performance of the conditional mean based predictions, we first follow the recent literature [Zhao et al. (2015), Kobayashi and Lambiotte (2016)] and use the Absolute Percentage Error (APE),

$$\text{APE} = \left| \frac{N(\tilde{T})_{\text{pred}} - N(\tilde{T})}{N(\tilde{T})} \right|,$$

to compare the accuracy of prediction by different models. Each prediction method under evaluation was applied to each of the 94,254 retweet cascades in the test data set with censoring times $T = 2, 4, \ldots, 12$ hours as most retweeting events would have occurred within the first few hours, based on our analysis on the training data set. For each censoring time, we calculated the APEs of the conditional mean predictions by the proposed model and by the two competing models. The predictions by the SEISMIC approach were calculated using the R package `seismic`. The predictions by the TiDeH model approach were calculated using the algorithm described in Kobayashi and Lambiotte (2016) with the window size parameter Δ_{obs} [cf. Kobayashi and Lambiotte (2016), page 194] in the estimation step set to one hour. Due to the lack of a principled approach to select the window size, we chose this value based on experimenting with different values and selecting the one that produced reasonable estimates of the infectivity function by visual inspection.

The boxplots of the APEs of conditional mean predictions by the three models (MaSEPTiDE, SEISMIC, and TiDeH) with different censoring times are shown in Figure 2. In each boxplot, the horizontal thick bar indicates the median APE, and the circular point indicates the mean APE. The actual values of the median and mean APE are given in Table 3. From Figure 2 and Table 3, the MaSEPTiDE prediction has consistently smaller median APE and mean APE for each T than the SEISMIC prediction. Compared to the TiDeH prediction, the MaSEPTiDE
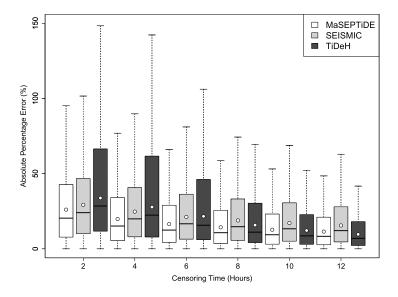
FIG. 2. *Boxplots of the Absolute Percentage Errors* (*APEs*) *of predictions by the MaSEPTiDE model, the SEISMIC and the TiDeH model, for censoring times $T = 2, 4, \ldots, 12$ hours. The horizontal thick bar in each boxplot indicates the median while the circular point indicates the respective mean of APEs. Both the mean and median of APEs demonstrate the superior performance and stability of the MaSEPTiDE model.*

prediction has clearly better performances when $T = 2, 4, 6$ hours, either by the median APE or by the mean APE. The performances of these two models are comparable when $T = 8$ hours, but the MaSEPTiDE is at a slight disadvantage when $T = 10, 12$ hours.

Despite its widespread use in evaluating prediction performance, the APE as a prediction error measure is not consistent with the feature of the predictive distribu-

TABLE 3
Median APEs and mean APEs of the popularity predictions by different approaches with observations up to various censoring times T
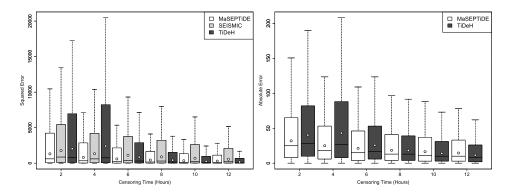
T (hours)	Median APE (%)			Mean APE (%)		
	MaSEPTiDE	SEISMIC	TiDeH	MaSEPTiDE	SEISMIC	TiDeH
2	19.1	22.8	23.7	26.1	29.3	33.8
4	13.9	18.6	17.1	19.8	24.8	27.8
6	11.2	15.1	12.7	16.5	21.2	21.6
8	9.5	13.1	9.3	14.3	18.9	15.8
10	8.2	11.7	7.4	12.7	17.1	12.2
12	7.3	10.6	5.9	11.4	15.5	9.6

FIG. 3.   *Left*: *squared prediction errors when the mean of the predictive distribution is used as the point prediction*. *Right*: *absolute prediction errors when the median is used*. *The thick horizontal bar in each boxplot shows the median of the errors*, *and the circular point shows the mean of the errors*.

tion used as a point prediction here, which is the expectation/mean; see Gneiting (2011) for a systematic discussion of this issue. A more appropriate prediction error measure when the conditional mean is used as the point prediction is the squared error. Therefore, we also calculated the squared errors of the predictions by the three models. The boxplots of the squared errors by the three models at different censoring times are shown in the left panel of Figure 3.

As the models can occasionally produce extremely large predictions, and even infinity in the case of SEISMIC, the outlying values were not shown in the boxplot for better visualization. The mean squared prediction errors (MSEs) at different censoring times are indicated by the circular points in the boxplots, and their squared roots, that is, the root mean squared errors (RMSEs), are shown in Table 4. From Figure 3 and Table 4, we note that, using the RMSE as the performance measure, the MaSEPTiDE prediction outperforms the SEISMIC prediction at all the censoring times, and similar to the conclusion drawn based on the median APEs,

TABLE 4
Root mean squared errors (*RMSEs*) *and mean absolute errors* (*MAEs*) *of predictions at different censoring times*

T (hours)	RMSE			MAE	
	MaSEPTiDE	SEISMIC	TiDeH	MaSEPTiDE	TiDeH
2	36.3	42.2	45.1	32.0	40.2
4	28.5	36.5	49.1	25.0	43.0
6	24.5	33.4	29.2	21.3	25.6
8	21.1	29.8	20.9	18.4	18.2
10	18.8	26.5	16.3	16.5	14.1
12	17.3	23.6	13.2	14.6	11.5

the MaSEPTiDE again, outperforms the TiDeH model when $T = 2, 4, 6$ hours but slightly underperforms when $T = 8, 10, 12$ hours. In comparison, the SEISMIC only outperforms the TiDeH model when $T = 2, 4$ hours.

To compare the prediction performances when using the conditional median by different models, we use the mean absolute error (MAE) as the criterion of comparison, as advised by Gneiting (2011). The conditional median prediction by the MaSEPTiDE model was calculated by the simulation based approach described in Section 2.5. The conditional median prediction by the TiDeH model was similarly calculated using a simulation based approach, although the simulation of the TiDeH model was achieved by a less efficient method where the events have to be simulated serially one after another, using the rejective method of Lewis and Shedler (1979). The conditional median prediction by the SEISMIC is not included in this comparison because this model does not specify the form of the intensity process beyond the censoring time, and therefore we cannot calculate the conditional median using the simulation based approach. The right panel of Figure 3 shows the absolute errors of the conditional median predictions by the MaSEPTiDE and the TiDeH models at different censoring times, where, as before, the circular points indicate the MAEs of the predictions at the corresponding censoring times. See also Table 4 for MAE values. By the MAE, the conditional median prediction by the MaSEPTiDE model is clearly superior to that by the TiDeH model at the censoring times $T = 2, 4, 6$ hours, and is comparable albeit slightly inferior at the larger censoring times $T = 8, 10, 12$ hours.

By all the performance evaluation criteria considered, the prediction by the MaSEPTiDE model is clearly more accurate than by the two competing models, especially when prediction needs to be made based on shorter observation times, for example, within six hours or shorter of the posting of the original tweet.

4. Conclusion and discussion.   In this work, we have proposed a marked self-exciting point process model, called the MaSEPTiDE, to model the retweeting dynamics and to predict tweet popularity. The MaSEPTiDE is able to model a large number of retweet cascades adequately, and its prediction performance is superior to those of the competing models and approaches in the literature that require the same input.

When the prediction is based on observing a cascade for a long period of time, the approach based on the TiDeH model by Kobayashi and Lambiotte (2016) is found to outperform our model by a small margin. However, considering the fact that this small advantage of the TiDeH model is not realized until the retweet cascades are observed for eight hours or longer, when the majority of the retweets would have already happened, its practical significance is rather limited. On the contrary, the approach based on the MaSEPTiDE is able to provide accurate predictions of the final popularity based on observations within two hours of posting of the original tweet. Another issue with the TiDeH model is that the nonparametric estimation step to obtain the initial raw estimate of the infectivity curve needs a large amount of data to work well. In fact, in their numerical experimentation,

Kobayashi and Lambiotte (2016) only verified the superior performance of their prediction approach relative to the SEISMIC on 738 very long cascades (containing 2,000 or more retweets), which account for less than 0.5% of all the cascades. In contrast, the approach based on the MaSEPTiDE is fully parametric and does not require as much data to estimate.

The specific parametric forms of the functions in the MaSEPTiDE model have been selected from a class of candidate models by comparing their goodness-of-fit on the retweet cascades in the training data set and identifying the model that can fit most of the cascades. In the class of candidate models, we have considered other parametric forms of the component functions, such as infectivity functions that decay at polynomial rate, and memory kernel functions that decay exponentially fast. The model with the specific forms of the component functions reported herein has the best goodness-of-fit on the training data.

To further improve the MaSEPTiDE model, more complex models, such as those that incorporate the calendar time effects [Fox et al. (2016), Kobayashi and Lambiotte (2016)] are worth considering. Another aspect of our approach that can be improved is that our approach still requires the observation of the retweet cascade for a substantial amount of time to accumulate enough data to identify the model, even though the required observation time is much less compared to approaches based on other models such as the TiDeH model. If we make stronger assumptions on the model parameters across the cascades, then parameter estimation might be achieved using only training data, which will allow us to predict its final popularity as soon as a tweet is published, or even before it is published.

Finally, an important limitation of the data considered in our work, originally collected by Zhao et al. (2015), is that the data contains only cascades with at least 49 retweets. Such data is by no means representative of all tweets published by Twitter users, as vast majority of the tweets do not get even a single retweet. Therefore, models developed based on such data are only useful for popularity predictions of reasonably popular tweets. To develop models suitable for the predictions of the popularity of average tweets, one would need to collect suitable random samples of tweets and their retweet cascades, and build models accordingly.

REFERENCES

AGARWAL, D., CHEN, B.-C. and ELANGO, P. (2009). Spatio-temporal models for estimating click-through rate. In *Proceedings of the* 18*th International Conference on World Wide Web* 21–30. ACM, New York.

AHMED, M., SPAGNA, S., HUICI, F. and NICCOLINI, S. (2013). A peek into the future: Predicting the evolution of popularity in user generated content. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* 607–616. ACM, New York.

ALVES, R. A., ASSUNÇÃO, R. and DE MELO, P. O. (2016). Burstiness scale: A highly parsimonious model for characterizing random series of events. Preprint. Availble at arXiv:1602.06431.

BAKSHY, E., HOFMAN, J. M., MASON, W. A. and WATTS, D. J. (2011). Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* 65–74. ACM, New York.

BARABASI, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435** 207–211.

CHA, M., HADDADI, H., BENEVENUTO, F. and GUMMADI, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the Fourth International Conference on Weblogs and Social Media* (*ICWSM*-2010) 10–17. AAAI Press, Palo Alto, CA.

CHEN, F. and HALL, P. (2013). Inference for a nonstationary self-exciting point process with an application in ultra-high frequency financial data modeling. *J. Appl. Probab.* **50** 1006–1024. MR3161370

CHEN, F. and HALL, P. (2016). Nonparametric estimation for self-exciting point processes—A parsimonious approach. *J. Comput. Graph. Statist.* **25** 209–224. MR3474044

CHEN, F. and STINDL, T. (2018). Direct likelihood evaluation for the renewal Hawkes process. *J. Comput. Graph. Statist.* **27** 119–131.

CRANE, R. and SORNETTE, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA* **105** 15649–15653.

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes*, *Vol. I*: *Elementary Theory and Methods*, 2nd ed. Springer, New York. MR1950431

FOX, E. W., SHORT, M. B., SCHOENBERG, F. P., CORONGES, K. D. and BERTOZZI, A. L. (2016). Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Amer. Statist. Assoc.* **111** 564–584. MR3538687

GAO, S., MA, J. and CHEN, Z. (2015). Modeling and predicting retweeting dynamics on microblogging platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* 107–116. ACM, New York.

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410

KOBAYASHI, R. and 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) 191–200. The AAAI Press, Palo Alto, CA.

KWAK, H., LEE, C., PARK, H. and MOON, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the* 19*th International Conference on World Wide Web* 591–600. ACM, New York.

LEWIS, P. A. W. and SHEDLER, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Nav. Res. Logist. Q.* **26** 403–413. MR0546120

LI, C.-T., SHAN, M.-K., JHENG, S.-H. and CHOU, K.-C. (2016). Exploiting concept drift to predict popularity of social multimedia in microblogs. *Inform. Sci.* **339** 310–331.

LYMPEROPOULOS, I. N. (2016). Predicting the popularity growth of online content: Model and algorithm. *Inform. Sci.* **369** 585–613.

MATSUBARA, Y., SAKURAI, Y., PRAKASH, B. A., LI, L. and FALOUTSOS, C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 6–14. ACM, New York.

MISHRA, S., RIZOIU, M.-A. and 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* 1069–1078. ACM, New York.

NAVEED, N., GOTTRON, T., KUNEGIS, J. and ALHADI, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on Twitter. In *Proceedings of the 3rd International Web Science Conference* Art. ID 8. ACM, New York.

NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Comput. J.* **7** 308–313. MR3363409

OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

SZABO, G. and HUBERMAN, B. A. (2010). Predicting the popularity of online content. *Commun. ACM* **53** 80–88.

TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G. and WELPE, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International Conference on Weblogs and Social Media* (*ICWSM*-2010) 178–185. AAAI Press, Palo Alto, CA.

WU, B., CHENG, W.-H., ZHANG, Y. and MEI, T. (2016). Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the* 2016 *ACM on Multimedia Conference* 1336–1344. ACM, New York.

YAN, Y., TAN, Z., GAO, X., TANG, S. and CHEN, G. (2016). STH-Bass: A spatial-temporal heterogeneous bass model to predict single-tweet popularity. In *International Conference on Database Systems for Advanced Applications* 18–32. Springer, Cham.

ZAMAN, T., FOX, E. B. and BRADLOW, E. T. (2014). A Bayesian approach for predicting the popularity of tweets. *Ann. Appl. Stat.* **8** 1583–1611. MR3271345

ZHAO, Q., ERDOGDU, M. A., HE, H. Y., RAJARAMAN, A. and LESKOVEC, J. (2015). SEISMIC: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1513–1522. ACM, New York.

SCHOOL OF MATHEMATICS AND STATISTICS
UNSW SYDNEY
SYDNEY NSW 2052
AUSTRALIA
E-MAIL: feng.chen@unsw.edu.au
          waihong.tan@student.unsw.edu.au