On the Choice of Functionals Obtained from the Predictive Distribution of Future Retweet Counts

Wai Hong Tan[1,a)] and Feng Chen[2,b)]

[1] *Universiti Malaysia Kelantan, Malaysia*
[2] *UNSW Sydney, Australia*

[a)] *Corresponding author: wai.hong@umk.edu.my*
[b)] *Electronic mail: feng.chen@unsw.edu.au*

Abstract. The prediction of future retweet counts for tweets shared on Twitter has been a topic of immense interest recently. Numerous models have been proposed for such prediction, with their accuracy being assessed using certain choices of evaluation metrics. Admittedly, the majority of predictive models involved have overlooked the problem on the use of theoretically optimal functionals as point predictions and resort to employing options which are more accessible like the predictive mean. This motivates our discussion wherein the practicality of using theoretically consistent functionals with respect to the evaluation metrics considered is put forth. We discuss how the median of order (–1) and harmonic median are optimal in theory relative to the mean and median absolute percentage errors respectively, followed by highlighting in contrast how predictive models extant in the literature may suggest otherwise. Specifically, using a Poisson model supported by a large corpus of Twitter data, our numerical experiments indicate that predictions based on different functionals derived from the predictive distribution do not vary materially across the different metrics used, although predictions stemming from the predictive mean are slightly yet consistently more accurate than those based on the other functionals. We further outline how consistent functionals can be obtained accordingly under the settings of more complex predictive models.

Keywords: harmonic median; importance sampling; order (–1) median; simulation; truncated distribution

INTRODUCTION

Information is disseminated and transferred throughout the Twitter network as retweet cascades. Let $\mathcal{R}$ denote the retweet process governing such information transfer, with time epochs $\{\tau_i\}$ distributed on the half-line $t > 0$ such that $\tau_i = \inf\{t > 0 : N(t) \geqslant i\}, i = 1, 2, \ldots$ and $N_{\mathcal{R}}(\cdot)$ be the associated counting process. Intuitively, the process $\mathcal{R}$ is initiated after the publication of a tweet. Conforming to the coordinated universal time (UTC), all tweets can be assumed to start at the relative time $\tau_0 = 0$, referred to as the ancestral tweets. The sequences of non-coinciding retweet times that follow are denoted correspondingly by $\tau_1 < \tau_2 < \ldots < \tau_{N_{\mathcal{R}}(T)}$ up to a certain censoring time T. For each retweet cascade, the d-dimensional features attached to the tweeter and retweeters are denoted respectively by κ_0^d and $\kappa_1^d, \kappa_2^d, \ldots, \kappa_{N_{\mathcal{R}}(T)}$.

The virality of a tweet, which corresponds to its travel speed throughout the Twitter network, is directly correlated to its future volume of retweets. The determinants of virality range from the more prominent features associated with tweeters such as their follower counts [1, 2] to instances like the tweet sentiments, the times and dates of tweet publications, the geolocations wherein the tweeters reside, and also features associated with the retweeters [3].

The problem of tweet popularity prediction has been gaining immense interest recently for its remarkable applicability, notably in online marketing [4, 5] and network dimensioning [6]. Extant literature has systematically discerned popularity prediction methods based on classes such as feature-driven and generative approaches [7] used primarily for pre- and post-publication predictions respectively, as well as micro- and macro-level approaches constructed based on the granularity of information [8].

When attempting to predict the future popularity levels of tweets, micro-level prediction methods are more insightful as they leverage the dynamics of heterogeneous users by giving them unique treatment, with indicative examples being the works of [9], [10] and [11]. The commonality shared between these models is that their formulations rely solely on $(\tau_i, m_i), i = 0, 1, 2, \ldots, N_{\mathcal{R}}(T)$, where $m \in \kappa$ is the number of followers, and that they are all post-publication prediction approaches motivated by the theory of point processes.

To overcome the limitation of these models whereby sufficient observations need to accumulate before a prediction can be made, [12] proposed an empirical Bayes method to estimate parameters and make predictions using knowledge internal and external to the tweets of interest. Specifically, pre-publication predictions can be made by depending on knowledge external to the target tweet, whilst post-publication predictions rely on both external and internal knowledge, with the dependency on the latter getting increasingly higher as T increases.

These aforesaid models share an important resemblance in that they are all based on fitting point processes to the observed retweet sequences up to a specific censoring time, and then projecting the fitted processes to a future time point. Different methodologies have been employed to fit these models, including nonparametric [9], least squares [10], maximum likelihood [11], and empirical Bayes [12] methods. In the next section, the consistency of prediction functionals relative to the evaluation metrics considered, together with the methods of acquisition for the functionals, will be discussed based on these estimated state-of-the-art models.

THE METRICS AND FUNCTIONALS

Models constructed based on the theory of point processes are powerful tools to forecast how retweet cascades will evolve in the future by capturing the dynamics of observations available, as demonstrated by the SEISMIC (self-exciting model of information cascades) of [9], the TiDeH (time-dependent Hawkes) of [10], the MaSEPTiDE (marked self-exciting process with time-dependent excitation function) of [11], and the EB (empirical Bayes) models of [12].

In order to obtain the predictive distribution for the number of retweets at a future time point $\tilde{T}$ by leveraging the dynamics up to time T, one ought to simulate the sample paths $N_{\mathcal{R}}$ within $(T, \tilde{T}]$, or $\tilde{N}_{\mathcal{R}}$ within $(0, \tilde{T} - T]$. Our discussion hereinafter focuses on models with finite-dimensional parameters where such simulations can be performed by using various approaches, therefore nonparametric models such as the SEISMIC is not directly applicable.

Based on sufficiently many simulation replications, the predictive distribution for the number of retweets implied by each of the estimated models can be obtained. Subsequently, the predictive distribution can be used to predict the future popularity levels through exploiting suitable functionals. For this purpose, let the predicted future popularity (irrespective of the functional chosen) be denoted by $N_{\mathcal{R}}(\tilde{T})^*$ and the actual future popularity be $N_{\mathcal{R}}(\tilde{T})$.

The performances of different prediction methods can be assessed using evaluation metrics like the root mean squared error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the median absolute percentage error (MdAPE), although in some instances different assessment methods have been used, such as those depending on the paired sample t-tests to compare the differences in the values of Akaike Information Criterion (AIC) for model selection [13]. The formulas to calculate the MSE and MAE are shown respectively in (1) and (2), while the MAPE and MdAPE can be obtained based on the mean and median of the values obtained from (3).

$$\text{MSE} = \sum_{j=1}^{n} (N_{\mathcal{R}}(\tilde{T})_j^* - N_{\mathcal{R}}(\tilde{T})_j)^2 / n \tag{1}$$

$$\text{MAE} = \sum_{j=1}^{n} |N_{\mathcal{R}}(\tilde{T})_j^* - N_{\mathcal{R}}(\tilde{T})_j| / n \tag{2}$$

$$\text{APE} = |(N_{\mathcal{R}}(\tilde{T})_j^* - N_{\mathcal{R}}(\tilde{T})_j) / N_{\mathcal{R}}(\tilde{T})_j| \times 100 \tag{3}$$

The root mean squared error, or the RMSE, as the name implies, is simply the squared root of (1). Note from (1), (2), and (3) that the subscript j denotes the j^{th} cascade, and n the total number of such cascades.

[14] asserted that different functionals are optimal based on different error metrics, and concluded that the RMSE and MAE are optimal relative to mean- and median-based predictions respectively. It is worth noting here that the RMSE will penalize larger errors more severely, and tends to get heavily distorted by the presence of outliers. This proves to be problematic for models such as the SEISMIC due to that under a propounded supercritical regime, the process will generate an infinite number of events. On another note, when the range of actual popularity values is known, measures based on the absolute errors, such as the MAE, will be relatively more useful.

The APE, or the absolute relative error expressed in percentage, is arguably a more informative metric to measure how much the predicted popularity deviates from the actual popularity value, and is particularly useful in comparing the efficiency of prediction methods when the popularity values are distributed very differently. In fact, both [9] and [10] used the MdAPE to evaluate the prediction performances of their proposed methodologies (using the conditional expectation or the predictive mean as the functional), since the median is known to be a robust estimator which is more resistant to the presence of outlying APE values. There is, however, a notable pitfall when using such median-based metric in assessing the performance of a prediction method, since it allows up to half of the predicted values to be arbitrarily bad. Therefore, statistical inferences and conclusions on how a prediction method outperforms the others should be made using both mean- and median-based metrics, which in this case are the MAPE and MdAPE.

Obtaining the conditional expectation and conditional median

The conditional expectation for the future popularity levels implied by state-of-the-art models can be obtained through various principled approaches. For instance, by relying on the branching process interpretation, the SEISMIC predicts the future popularity levels of tweets through incorporating scaling constants which account for the gradual decay of tweet virality and possible mutuality in the pool of followers. As for the TiDeH and the MaSEPTiDE models, predictions using the conditional expectation can be obtained through solving self-consistent integral equations.

The conditional median, on another note, can be obtained via simulation-based approaches. For the SEISMIC, however, since it does not specify the form of its intensity process beyond the censoring time, we cannot directly apply the simulation-based approach to obtain its conditional median. The conditional median predictions by the TiDeH model, notably, can be calculated using the predictive distribution with events being simulated serially one after another, using the rejective method of [15]. The conditional median predictions by the MaSEPTiDE model can similarly be calculated by the simulation-based approach, such as that motivated by the cascading algorithm of [16].

For more general models, such as those with events arriving according to inhomogeneous Poisson processes, as in the empirically-motivated Poisson model presented by [12], the conditional expectation and median are relatively easy to obtain. Specifically, the Poisson processes are simulated using a time-varying intensity function for a sufficient number of times to obtain the predictive distribution where the conditional expectation and median can be extracted.

Although the predictive mean and median are not optimal relative to the MAPE or MdAPE in general, they are widely used in popularity predictions even when the MAPE or MdAPE is used as the metric. This is due to that they are approximately optimal when the predictive distribution is unimodal, and are typically much easier to obtain than functionals optimal relative to either of these metrics, aptly termed as the order (-1) median and harmonic median respectively, detailed as follows.

Obtaining the order (–1) median and harmonic median

We have discussed that the RMSE and MAE are metrics consistent with the predictive mean and median respectively. Besides these metrics, the MAPE and MdAPE are frequently used in assessing the accuracy of tweet popularity prediction methods in the literature, which are theoretically inconsistent with the predictive mean and median. Thus, our discussion herein focuses on the optimal functionals for these two metrics, with special emphasis on how the functionals can be obtained for different models proposed in the literature.

For the EB Poisson model

We make assumption on that the predictive distribution F is supported by positive reals $\mathbb{R}^+$. Then, as implied by [14], the point prediction that is optimal relative to the MAPE is the *order (–1) median* of F, denoted by $\text{med}^{(-1)}(F)$, and defined as the median of the tilted distribution,

$$\left(\int_0^\infty y^{-1}\,\mathrm{d}F(y)\right)^{-1} y^{-1}\,\mathrm{d}F(y).$$

Here, we note that $\text{med}^{(-1)}(F)$ is defined only when $\int_0^\infty y^{-1}\,\mathrm{d}F(y) < \infty$, which is clearly true in the case of empirically-motivated Poisson model since the predictive distribution has a lower bound of $\underline{r} = 49$, attributable to how the data was collected where only tweets that are relatively popular observed over the course of $\tilde{T} = 7$ days are considered.

In general, the computations of the order (-1) median and the harmonic median require numerical procedures. A general Monte Carlo approach to compute $\text{med}^{(-1)}(F)$ is to use *importance sampling* with the following procedures,

1. Simulate a large independent and identically distributed (i.i.d) sample $\mathcal{S} = \{y_i, i = 1, 2, \ldots, B\}$ from the distribution F.

2. Take a bootstrap resample $\mathcal{S}^* = \{y_i^*, i = 1, 2, \ldots, B\}$ from the simulated sample $\mathcal{S}$ (with replacement) where the selection probabilities for y_i are proportional to y_i^{-1}.

3. Approximate $\text{med}^{(-1)}(F)$ by the median of $\mathcal{S}^*$.

Under the Poisson process model considered in [12], F is a truncated and shifted Poisson distribution. Therefore, to simulate from F, we can first simulate from the truncated Poisson distribution, with the lower bound $\max\{\underline{r}-N_{\mathcal{R}}(T),0\}$, using either the rejective method or the inversion method, and then add $N_{\mathcal{R}}(T)$ to the simulated values. The choice of the method here depends on the value of the lower bound relative to the mean of the (untruncated) Poisson distribution. In particular, when the lower bound is smaller than the Poissonian mean, then the rejective method is more efficient. On the contrary, when the Poissonian mean is much smaller than the lower bound, then the inversion method is more efficient.

The point prediction that is optimal relative to the MdAPE, on the other hand, can be shown to be the harmonic mean of the two closest numbers $l \leqslant u$ such that $F(u)-F(l-) \geqslant 1/2$ and $F(u-)-F(l) \leqslant 1/2$, which we refer to as the *harmonic median*, and is conveniently denoted by $\mathrm{hamed}(F)$. Note, when F is continuous, the constraints on l and u in the definition of $\mathrm{hamed}(F)$ can be simplified to $F(u)-F(l)=1/2$.

For a general predictive distribution F, the computation of its harmonic median can be challenging. However, for the truncated and shifted Poisson distribution under the Poisson process model that we are dealing with here, the numerical computation is relatively easy. First, note that when the mode of the Poisson distribution is $\underline{r}-N_{\mathcal{R}}(T)$ at max, the probability function of the truncated Poisson distribution is a decreasing function, therefore $l=\underline{r}$, $u = N_{\mathcal{R}}(T)+\mathrm{med}[N_{\mathcal{R}}(\tilde{T})-N_{\mathcal{R}}(T) | N_{\mathcal{R}}(\tilde{T})-N_{\mathcal{R}}(T) \geqslant \underline{r}-N_{\mathcal{R}}(T)]$, and

$$\mathrm{hamed}(F) = 2/(l^{-1}+u^{-1}).$$

In contrast, when the mode of the Poisson distribution is greater than $\underline{r}-N_{\mathcal{R}}(T)$, we can calculate the harmonic median based on the following algorithm, where $f(\cdot)$ denotes the probability mass function of $N_{\mathcal{R}}(\tilde{T})-N_{\mathcal{R}}(T)$ given that it is at least $\underline{r}-N_{\mathcal{R}}(T)$,

1. Set both l and u to the mode of the Poisson distribution, and if there are two modes, set l to the smaller mode and u to the larger one.

2. While $\sum_{i:l<i\leqslant u} f(i) < 1/2$ is true, repeat the following:

   If $l > \max\{\underline{r}-N_{\mathcal{R}}(T),0\}$ and $f(l-1) > f(u+1)$, set $l \leftarrow l-1$;
   otherwise, set $u \leftarrow u+1$.

3. Set $l \leftarrow l+N_{\mathcal{R}}(T)$ and $u \leftarrow u+N_{\mathcal{R}}(T)$.

4. Return $2/(l^{-1}+u^{-1})$ as the harmonic median.

For other models

We consider herein cases when the order (-1) and harmonic medians are of interest for other models, in particular the TiDeH and the MaSEPTiDE models, by referring to the general procedures outlined above.

In obtaining the order (-1) median, we note for point process models such as the MaSEPTiDE or the TiDeH model that the truncated distribution for $N_{\mathcal{R}}(\tilde{T})-N_{\mathcal{R}}(T)$ might not have an explicit or otherwise easy-to-compute density or mass function, and therefore the inversion sampling method is not applicable. Under such a circumstance, we can use the rejective method, that is, by simulating values from the untruncated distribution and retaining only values with the condition of being at least $\underline{r}-N_{\mathcal{R}}(T)$. However, a potential issue with this method is that none of the values simulated from the untruncated distribution meets the retention condition, despite a large number of values have been simulated. When this happens, we can simply approximate the order (-1) median of the distribution by the corresponding $\underline{r}$.

For harmonic median, under models where the probability mass function of $N_{\mathcal{R}}(\tilde{T})-N_{\mathcal{R}}(T)$ is not available but it is relatively easy to simulate from the distribution, we can try to get a sample from the truncated distribution using the rejective method, and then use the empirical mass function in the algorithm highlighted before to obtain an estimate of the harmonic median. Under the scenario when a truncated sample is extremely difficult to acquire, we can again approximate the desired harmonic median by the corresponding lower bound $\underline{r}$.

Essentially, the practice highlighted here can also be applied on other point process models such as those of [17], [18], and [19] to get their corresponding order (-1) and harmonic medians based on their simulated predictive distributions. In summary, prediction functionals which are (theoretically) optimal relative to their respective evaluation metrics are exhibited in Table I. It should be noted from Table I that when using certain choices of metrics, one must not overlook the problem on the theoretical optimality of the corresponding functionals, to prevent the possibility of making misguided inferences.

TABLE I. Prediction functionals consistent with the target error metrics.

Target metric	Consistent functional
RMSE	Mean
MAE	Median
MAPE	Order (-1) median
MdAPE	Harmonic median

NUMERICAL RESULTS

The results shown in this section shall be based on the empirically-motivated Poisson process model discussed in the previous section, for its convenient distributional properties and ease of interpretation. It is worth noting here that the EB estimation approach is essentially a penalized maximum likelihood approach whereby the likelihood function has a larger curvature and is easier to maximize than its unpenalized counterpart. In addition, the prior distribution for the parameters under the approach is inspired by the concept of confidence distribution, which gives a regularization effect on the maximum likelihood estimates, and enables prediction at time zero.

Using the empirical Bayes (EB) approach highlighted by [12], the EB estimates for the Poisson model for each of the 94,254 retweet cascades in the test data [20] at $T = 0, 2, 4, \ldots, 12$ hours were obtained. Following this, the future popularity levels for these tweets at $\tilde{T} = 7$ days were predicted using the mean, median, median of order (-1), and harmonic median of the predictive distributions under the specified censoring times. The corresponding RMSE, MAE, MAPE, and MdAPE of predictions were also calculated.

The prediction accuracy of different functionals according to different metrics at censoring time zero based on the estimated EB Poisson model is shown in Table II, and the results of comparisons at later censoring times ($T = 2, 4, \ldots, 12$ hours) are shown in Table III. By any of the four metrics, the point predictions by the four functionals

TABLE II. The accuracy of different functionals at censoring time zero, using the complete test data. Point predictions based on the predictive mean are consistently more accurate than those based on the other functionals.

	RMSE	MAE	MAPE	MdAPE
Mean	382.47	135.67	47.86%	43.57%
Median	382.57	135.94	48.08%	43.86%
Order (-1) median	382.82	136.23	48.07%	43.96%
Harmonic median	382.86	136.56	48.49%	44.47%

TABLE III. The accuracy of different prediction functionals at censoring times $T = 2, 4, \ldots, 12$ hours, using the complete test data. Point predictions based on the predictive mean seem consistently more accurate than those based on the other functionals.

$T = 2$ hours	RMSE	MAE	MAPE	MdAPE	$T = 8$ hours	RMSE	MAE	MAPE	MdAPE
Mean	390.81	69.73	23.35%	16.54%	Mean	349.90	53.12	15.94%	8.03%
Median	390.83	69.86	23.48%	16.73%	Median	349.91	53.19	16.01%	8.12%
Order (–1) median	390.82	69.90	23.51%	16.81%	Order (–1) median	349.87	53.19	16.01%	8.12%
Harmonic median	390.85	70.03	23.62%	16.95%	Harmonic median	349.89	53.27	16.08%	8.23%

$T = 4$ hours	RMSE	MAE	MAPE	MdAPE	$T = 10$ hours	RMSE	MAE	MAPE	MdAPE
Mean	303.99	60.12	19.50%	11.74%	Mean	391.89	51.23	14.84%	6.98%
Median	304.01	60.22	19.61%	11.93%	Median	391.89	51.29	14.90%	7.14%
Order (–1) median	303.98	60.24	19.62%	11.94%	Order (–1) median	391.86	51.28	14.89%	7.14%
Harmonic median	304.01	60.34	19.72%	12.09%	Harmonic median	391.88	51.35	14.96%	7.20%

$T = 6$ hours	RMSE	MAE	MAPE	MdAPE	$T = 12$ hours	RMSE	MAE	MAPE	MdAPE
Mean	303.05	55.29	17.37%	9.40%	Mean	405.18	49.05	13.86%	6.18%
Median	303.06	55.37	17.46%	9.52%	Median	405.18	49.10	13.91%	6.25%
Order (–1) median	303.02	55.37	17.46%	9.52%	Order (–1) median	405.14	49.08	13.91%	6.25%
Harmonic median	303.05	55.47	17.55%	9.67%	Harmonic median	405.16	49.15	13.96%	6.35%

seem to have comparable accuracy, although the predictions by the predictive mean are slightly yet consistently more accurate than those based on the other functionals. In addition, our numerical results indicate that the predictions using different functionals of the predictive distributions across the times are nearly identical to each other, with the maximum absolute difference being merely 9.91.

Surprisingly, the theoretically optimal functional for a specific error metric does not lead to a more accurate prediction by the corresponding metric, although in our simulations the optimal functionals do produce slightly more accurate predictions than the other functionals, by the compatible error metrics. This may be due to that the numbers of retweets may not follow the Poisson distributions exactly while they do in the simulations. Based on this observation and the ease of computation of the mean of the predictive distribution, it appears sensible to use the predictive mean as the point prediction, irrespective of the error metric chosen.

Moreover, based on Table II and Table III we can see that the MAE, MAPE and MdAPE all exhibit increasingly better performances with larger censoring times, but the RMSE seems to fluctuate indefinitely. This issue stems from the presence of grossly erroneous predictions from which their errors are severely magnified by the metric, which then conceal the good performance of the model. Thus, our recommendation regarding the use of unitless error measures such as those based on the APE is further substantiated.

On another remark, Table IV shows the actual popularity values of retweet cascades in both the training and test data sets. It can be observed from the table that the tweet popularity levels are highly heterogeneous, which suggests

TABLE IV. The summary statistics for the actual popularity values at $\tilde{T} = 7$ days, with Q_1, Q_2 and Q_3 denoting the 1^{st}, 2^{nd}, and 3^{rd} quartiles respectively.

	Min	Q_1	Q_2	Q_3	Max	Mean
Training	49	70	109	216	33484	205.5
Test	49	70	110	222	17183	210.7

that the APE is comparably more informative and suitable in assessing the accuracy of tweet popularity predictions.

CONCLUSION

This paper aims to testify if using theoretically optimal functionals relative to the error metrics chosen is at all practical in the context of tweet popularity prediction. By convention, the accuracy of tweet-oriented predictive models is assessed using metrics like the RMSE, MAE, MAPE, or MdAPE, with the archetypal functional being the predictive mean or the conditional expectation. The prime choice of functional has been the predictive mean in extant literature due to its ease of acquisition, despite the prevailing issue on its inconsistency with different metrics.

In term of theoretical optimality, it has been demonstrated that the predictive mean and median are functionals consistent with the squared and absolute errors respectively, making the RMSE and MAE natural choices of error metrics in assessing the accuracy of predictive models when these features of the predictive distribution are available. However, due to the observation that tweets are highly heterogeneous in popularity levels, it appears sensible to opt for the MAPE and MdAPE when assessing the performances of different tweet popularity prediction methods.

Although the consistent prediction functionals based on the MAPE and MdAPE are the order (-1) and harmonic medians respectively, the use of predictive mean as the functional is justifiable. First, the predictive mean is relatively easier to obtain compared the two proposed functionals with considerable computational complexity. Second, based on our numerical experiments (on the empirically-motivated Poisson model), predictions based on different functionals do not differ materially by the different metrics used, with the predictions based on the predictive mean being slightly yet consistently more accurate than those based on the other functionals.

As a concluding remark, when the numbers of retweets are generated through simulations, the optimal functionals do produce more accurate predictions than the other functionals, by the compatible error metrics. In contrast, when real tweet data sets are used, the theoretically optimal functional for a specific error metric does not necessarily lead to a more accurate prediction by the corresponding metric, due to that the numbers of retweets may not follow the specified distribution exactly. The use of theoretically optimal prediction functionals relative to the assessment metrics considered under settings outside the framework of retweet distributions is worth exploring in future work.

REFERENCES

1. K. Lerman and T. Hogg, "Using a model of social dynamics to predict popularity of news," in *Proceedings of the 19th international conference on World wide web* (2010) pp. 621–630.
2. S.-A. A. Jin and J. Phua, "Following celebrities' tweets about brands: The impact of twitter-based electronic word-of-mouth on consumers' source credibility perception, buying intention, and social identification with celebrities," Journal of advertising **43**, 181–195 (2014).
3. M. Farajtabar, S. Yousefi, L. Q. Tran, L. Song, and H. Zha, "A continuous-time mutually-exciting point process framework for prioritizing events in social media," arXiv preprint arXiv:1511.04145 (2015).
4. H. Lakkaraju and J. Ajmera, "Attention prediction on social media brand pages," in *Proceedings of the 20th ACM international conference on Information and knowledge management* (ACM, 2011) pp. 2157–2160.
5. A. Kupavskii, A. Umnov, G. Gusev, and P. Serdyukov, "Predicting the audience size of a tweet," in *ICWSM* (2013).
6. B. Han, P. Hui, V. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," IEEE Transactions on Mobile Computing **11**, 821–834 (2012).
7. S. Mishra, M.-A. Rizoiu, and L. Xie, "Feature driven and point process approaches for popularity prediction," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (ACM, 2016) pp. 1069–1078.
8. A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," Journal of Internet Services and Applications **5**, 8 (2014).
9. Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, "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, 2015) pp. 1513–1522.
10. R. Kobayashi and R. Lambiotte, "TiDeH: Time-dependent Hawkes process for predicting retweet dynamics," in *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)* (Association for the Advancement of Artificial Intelligence, 2016) pp. 191–200.
11. F. Chen and W. H. Tan, "Marked self-exciting point process modelling of information diffusion on Twitter," Ann. Appl. Statist. **12**, 2175–2196 (2018).
12. W. H. Tan and F. Chen, "Predicting the popularity of tweets using internal and external knowledge: an empirical bayes type approach," AStA Advances in Statistical Analysis , 1–18 (2021).
13. I. C. Dezza, J. Y. Angela, A. Cleeremans, and W. Alexander, "Learning the value of information and reward over time when solving exploration-exploitation problems," Scientific reports **7**, 1–13 (2017).
14. T. Gneiting, "Making and evaluating point forecasts," Journal of the American Statistical Association **106**, 746–762 (2011).
15. P. A. Lewis and G. S. Shedler, "Simulation of nonhomogeneous Poisson processes by thinning," Naval Research Logistics Quarterly **26**, 403–413 (1979).
16. F. Chen and P. Hall, "Nonparametric estimation for self-exciting point processes - a parsimonious approach," Journal of Computational and Graphical Statistics **25**, 209–224 (2016).
17. A. H. Zadeh and R. Sharda, "Modeling brand post popularity dynamics in online social networks," Decision Support Systems **65**, 59–68 (2014).
18. M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, "Expecting to be hip: Hawkes intensity processes for social media popularity," in *Proceedings of the 26th International Conference on World Wide Web* (2017) pp. 735–744.
19. E. Yoo, B. Gu, and E. Rabinovich, "Diffusion on social media platforms: A point process model for interaction among similar content," Journal of Management Information Systems **36**, 1105–1141 (2019).
20. Available from http://snap.stanford.edu/seismic/.