

# Modelling structure and predicting dynamics of discussion threads in online boards

ALEXEY N. MEDVEDEV\*,

NaXys, Université de Namur, 5000 Namur, Belgium

ICTEAM, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

\*Corresponding author: [an\\_medvedev@yahoo.com](mailto:an_medvedev@yahoo.com)

JEAN-CHARLES DELVENNE

ICTEAM and CORE, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium

[jean-charles.delvenne@uclouvain.be](mailto:jean-charles.delvenne@uclouvain.be)

AND

RENAUD LAMBIOTTE

NaXys, Université de Namur, 5000 Namur, Belgium

Mathematical Institute, University of Oxford, OX2 6GG Oxford, UK

[renaud.lambiotte@maths.ox.ac.uk](mailto:renaud.lambiotte@maths.ox.ac.uk)

## Abstract

Internet boards are platforms for online discussions about a variety of topics. On these boards, individuals may start a new thread on a specific matter, or leave comments in an existing discussion. The resulting collective process leads to the formation of ‘discussion trees’, where nodes represent a post and comments, and an edge represents a ‘reply-to’ relation. The structure of discussion trees has been analysed in previous works, but only from a static perspective. In this paper, we focus on their structural and dynamical properties by modelling their formation as a self-exciting Hawkes process. We first study a Reddit dataset to show that the structure of the trees resemble those produced by a Galton-Watson process with a special root offspring distribution. The dynamical aspect of the model is then used to predict future commenting activity and the final size of a discussion tree. We compare the efficiency of our approach with previous works and show its superiority for the prediction of the dynamics of discussion.

**Keywords:** complex networks, temporal networks, Hawkes processes, bursty time series, cascade prediction

2000 Math Subject Classification: 90B18, 60K35, 60G55, 82C99

# 1 Introduction

Online social media offer a rich source of information for the study of social behaviour. Depending on the platform, different types of tree-like cascading patterns emerge as a consequence of social interactions [1, 11]. For example, on Twitter or on Facebook, people interact via sharing or retweeting content, which may lead to large cascades of events [18, 6, 33]. In email networks, people may forward messages to their acquaintances, resulting in cascading trees of email forwards [14]. In online boards like Digg or Reddit, people interact by discussing certain posts, which leads to the formation of reply cascades or, so-called, ‘discussion trees’ [17]. In general, researchers have been interested in two main questions: what is the shape of these cascades and what is the dynamics of their evolution?

Several works have focused on the structural properties of cascades and on the mechanisms that could generate them. In [9], for instance, the authors considered cascades in four large Internet boards, including Reddit, and proposed a generating model based on preferential attachment mechanism. In [10], the authors enriched this model by incorporating a notion of attractiveness to the comments, and aimed at reproducing the relation between the width and the depth of discussion trees. Note that these models focus exclusively on the structure of the trees and not on the timings at which events take place. In [30], in contrast, the authors introduce a theoretical model for the structural and temporal evolution of discussions, based on the notion of Lévy processes, but the mean-field nature of the model limits its calibration with real-world datasets, and thus its applicability. Up to our knowledge, the problem of predicting both structure and dynamics of discussions remains open.

The interplay between the structure and the dynamics of cascades has been considered in other social media data sets, for instance as a feature in a machine learning framework to predict successful cascades from unsuccessful ones [3]. From a modelling perspective, word-of-mouth cascades in email networks have been modelled by Bellman-Harris branching processes with lognormal inter-branching times [14]. Hawkes processes were successfully applied to study and predict the evolution of retweet cascades in Twitter [33, 15]. The structure of so-called ‘reply trees’ was also studied by Nishi et. al in [21], where authors proposed simple principles for the formation of such trees. It is important to stress that the mechanisms behind retweet cascades and reply trees on one side, and discussion trees on the other side, are intrinsically different, due to the different designs of their platforms. The former is strongly affected by the structure of an underlying social network, for instance the followee-follower network of Twitter, while the latter makes content equally available to any visitor of a website, principally based on recency.

In this paper, we concentrate on ‘discussion trees’ formed for each post in Reddit, one of the largest online boards, and analyse the dataset of all comments to all posts initiated in the period between Jan 2008 and Jan 2015. Our main contribution is the introduction of a model for the growth of discussion trees based on Hawkes processes and its validation on the dataset. The model incorporates both structure and dynamics, and is used for predicting the future commenting activity. The paper is organized as follows. In Section 2, we present and describe

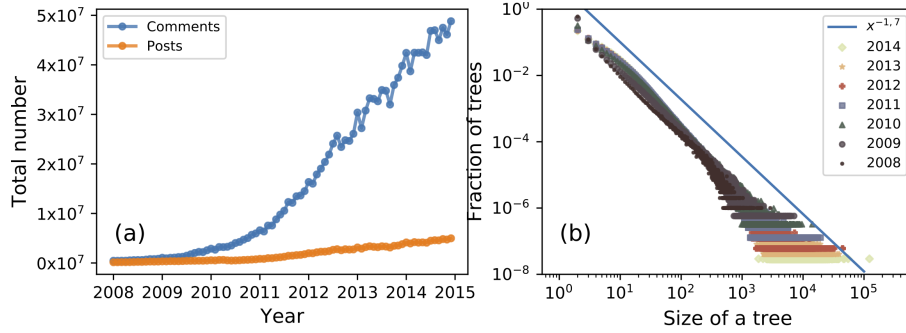


Figure 1: (a) Total monthly submissions to Reddit: orange curve represents posts and blue curve represents comments. (b) Distribution of tree sizes, collected for each year from 2008 to 2014. The power-law  $\sim x^{-1.7}$ , in blue, is a guide for an eye.

the main properties of the dataset, in Section 3 we present the model with parameters estimation procedure, in Section 4 we set up the evaluation procedure and present the existing models that we use as baselines for comparison, in Section 5 we present the results and we conclude the paper with the discussion in Section 6.

## 2 Dataset

Reddit is organised as follows: users post a message on a particular subject, which appears on the website and becomes available to all users. Each post has a section for comments, where users write a reply to the post or to previous comments of this post. In the following paper, we analyze the dataset of all posts and comments submitted to Reddit between Jan, 2008 and Jan, 2015 [29, 28]. Comments form a discussion, which can be represented as a *rooted tree*, where the root is a designated node representing the post itself and each other node represents a comment. There is a link between two nodes if there is a ‘reply-to’ relation between them. We disregard the edge direction as it may be recovered from the timestamps or tree level distance from the root. Each post and comment contains a unique id and information about its author, its content, the creation timestamp and a link to the node to which it replies. In the following, we disregard the information on the content of posts and comments, and concentrate mainly on the temporal and structural properties of the trees.

The dataset in total contains more than 150 million posts and around 1.4 billion comments. The stable growth of the number of total monthly submissions is presented on Figure 1, (a). The growth rate of both posts and comments increases between 2010 and 2012, and remains roughly stable afterwards, which is probably related to the release of a mobile version of the website for smartphones [23]. While the total number of discussion trees grows in time, their size distribution remains approximately the same (see Figure 1, (b)), with a shape close to a power-law  $\mathbb{P}(\xi > x) \sim x^{-\alpha}$  with  $\alpha = 1.7$ .

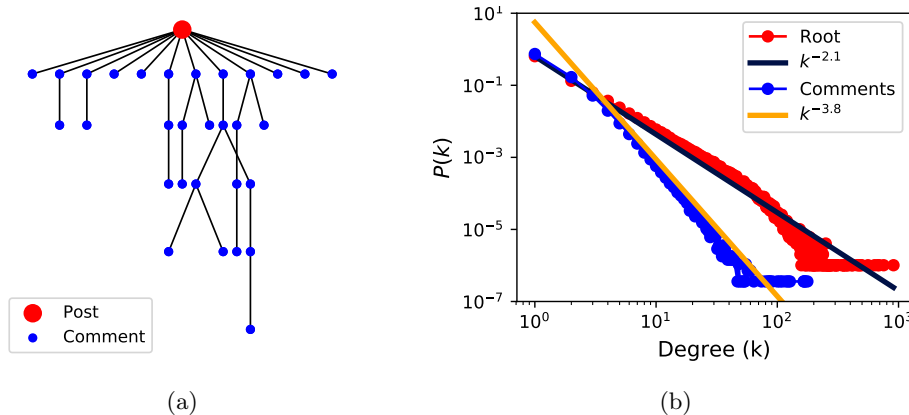


Figure 2: (a) Example of a discussion tree. The root is coloured in red and depicted with a larger size. (b) Distribution of degrees for the roots and for comments in 2008. Roots have a significantly larger degree, on average, than comments.

The Reddit platform is designed such that users are first exposed to the content and general information of a post, while comments are only visible afterwards. This difference implies that the root has a larger degree, on average, than a comment, as shown in Figure 2, (a), an observation already made in other Internet boards [9] and for reply trees in Twitter [21]. This observation has important consequences, as it suggests that the degree of the root is determined by a different process than for other nodes. This conclusion is confirmed in Figure 2, (b), where we show the degree distribution of the root (in red) and the forward degree distribution of all other vertices (in blue) in 2008 are clearly different. By definition, the *forward degree* is the degree of a comment minus one, and it corresponds to the number of comments made in reply to a certain comment. Both distributions are well approximated by a power-law with exponents close to 2.1 for the root and 3.8 for the comments. Note that the forward degree distribution does not appear to change with the distance to the root. Similar distributions were obtained for other years.

Another argument supporting the evidence of the different nature of comments and posts can be built from the temporal characteristics of the trees. Consider the discussion tree  $\mathcal{T}$  with node set  $V = \{v_1, \dots, v_n\}$ , where  $v_1 = s$  is the root. Denote as  $\tau_{v_i} \geq 0$  the timestamp of the node  $v_i$ , where  $1 \leq i \leq n$ . We further assume that  $\tau_{v_1} = 0$ , as it is possible to make an offset by the posting time. Consider a node  $v$  and its forward neighbour  $u$ . We call the difference  $\tau_v - \tau_u$  the *response time* of the node  $v$  to the node  $u$ , namely the node to which  $v$  replies to. The data shows that the response time distribution follows a different pattern for the root and for the comments (see Figure 3). In particular, the distribution for the root within 36 hours shows an initial power-law increase followed by an exponential decay (see inset to Figure 3, (a)), which can be well fitted by a pdf of the Weibull distribution:

$$\text{Weib}(a, b, \alpha) = a(\alpha/b) (t/b)^{\alpha-1} \exp(-(t/b)^\alpha),$$

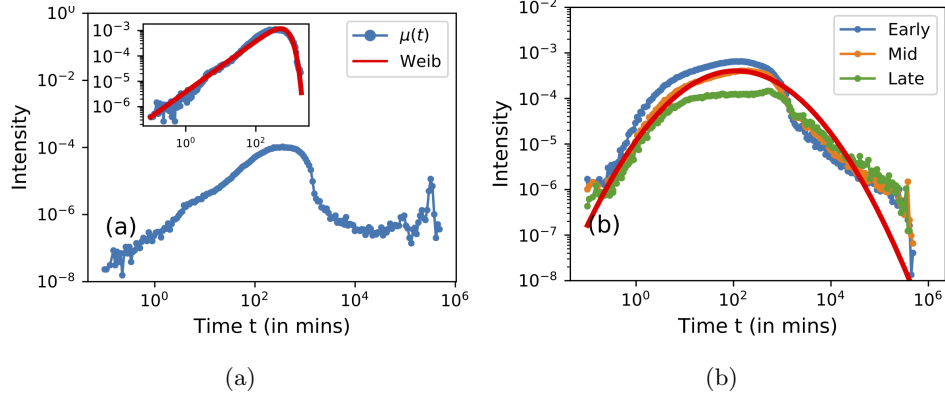


Figure 3: Response time distributions for (a) the root and (b) comments. The inset of (a) depicts the intensity of response times for the roots within 36 hours, along with the Weibull pdf fit given with the red curve. The response times for comments on (b) are divided into three sets: *early* comments that appeared within 6 hours from post's appearance, *mid* comments, created within 6 and 24 hours after the post's creation and the remaining *late* comments. The red line corresponds to a lognormal distribution.

where  $a > 0$ ,  $b > 0$ ,  $\alpha > 0$ . Weibull distribution is used in survival analysis and extreme value theory, as it represents a limit distribution of a minimum of  $n$  i.i.d. random variables, satisfying a specific criteria [13]. We observe from the Figure 3, (a), that the tail of the response time distribution does not follow any specific pattern, which may be the consequence of aggregation of many posts with different parameters  $a$  and  $b$ .

In principle, one could have expected that the age of a comment affects how many new comments it generates, and when those are generated. Thus we divide the comments into three subsets: 1) *early comments* that appeared within 6 hours, 2) *mid comments*, created within 6 and 24 hours and 3) the remaining *late comments*. According to the division we have about 1.5% of total comments classified as late and about 12% as mid. In each case, as we notice on Figure 3, (b), due to the absence of abrupt exponential decay for large times, the response time distribution is well fitted by a lognormal distribution:

$$\text{LogN}(\mu, \sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right),$$

where  $\mu, \sigma > 0$ . The lognormal distribution naturally emerges when considering the product of i.i.d. random variables with finite variance (see e.g. [7]) and it has been used in a variety of problems, for instance as a kernel to model citation dynamics [31, 26]. One could also hypothesize that the level of the comment may have an impact onto response time, but we did not find any such significant influence in the data.

To get a further insight into the temporal characteristics of the dataset, we consider, for each discussion, the time series  $\tau$  of creation times of each comment, defined as an ordered ascending collection of timings  $\tau = \{\tau_1 < \tau_2 < \dots < \tau_n\}$ . Due to its non-stationarity, we characterise

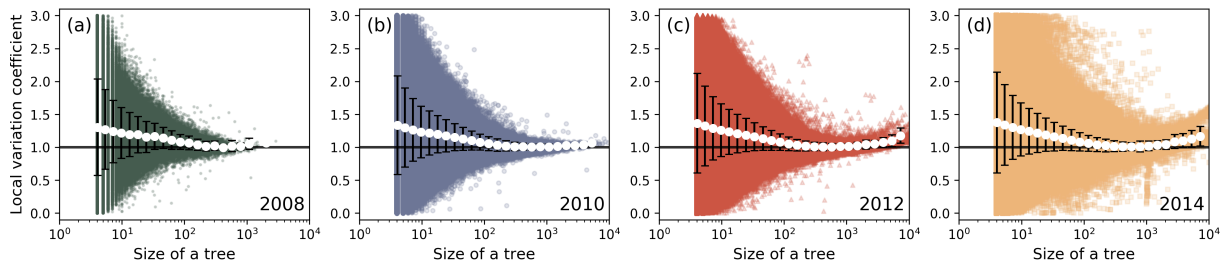


Figure 4:  $LV$  scores for discussion trees in the dataset for : (a) 2008, (b) 2010, (c) 2012, (d) 2014. The scatter plot of individual scores is accompanied with the mean score in white and bars for the standard deviation. It is observed that the size of trees increases in time, however the overall pattern of the distribution of  $LV$  coefficients remains similar. For each year, we present the tree sizes in the range from 1 to 10000 nodes, and as the tree size grows, the  $LV$  coefficients tend to be closer 1 on average (given as white circles) and show less dispersion as well (bars represent standard deviation).

correlations in the time series by its *local variation coefficient*  $LV$  [27, 16, 20, 22, 25]. The local variation coefficient is defined by comparing temporal variations with their local rates

$$LV(\tau) = \frac{3}{n-1} \sum_{i=1}^{n-1} \left( \frac{\delta\tau_{i+1} - \delta\tau_i}{\delta\tau_{i+1} + \delta\tau_i} \right)^2,$$

where  $\delta\tau_i = \tau_i - \tau_{i-1}$ , and is thus specifically designed for non-stationary processes. The coefficient takes values in the interval  $(0, 3)$ .  $LV(\tau)$  is equal to 1 when the point process that generates  $\tau$  is an inhomogeneous Poisson process. Deviations from 1 originate from local correlations in the underlying time series, either under the form of pairwise correlations between successive inter-event time intervals, e.g.  $\delta\tau_{i+1}$  and  $\delta\tau_i$ , which tend to decrease  $LV$ , or because the inter-event time distribution is non-exponential [25]. In Figure 4, we observe a relation between the size of a discussion tree and its value of  $LV$ , as larger trees tend to show values closer to those of Poisson process. We will provide an explanation for this observation in Section 5.

### 3 Hawkes model of discussion trees

A discussion tree may grow in two ways, either by adding comments to the original post or by adding a comment to an existing comment. As we showed, the temporal and structural properties of each mechanism appears to be different. We incorporate these findings in a self-exciting Hawkes process, a well-known point process where the intensity (i.e. rate function) evolves as

$$\lambda(t) = \mu(t) + n_b \sum_{i: \tau_i < t} \phi(t - \tau_i),$$

and where  $\mu(t)$  is the background intensity,  $n_b$  is the branching number and  $\phi(t)$  is the memory kernel. The process is self-exciting because the rate increases after the realisation of an event  $i$

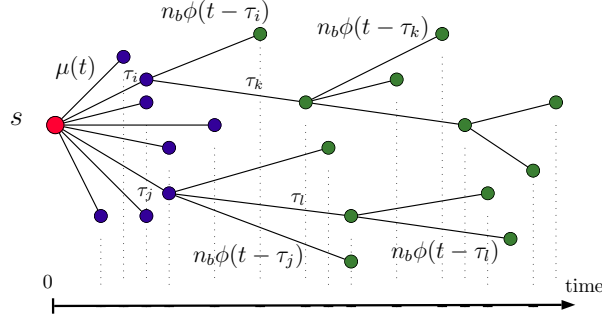


Figure 5: Graphical description of a Hawkes process as a branching process.

at time  $\tau_i$  with a kernel  $\phi(t - \tau_i)$  [12]. By definition, the memory kernel satisfies the condition that  $\langle \phi \rangle = \int_0^\infty \phi(t) dt = 1$  and it can be interpreted as the probability that a new triggered event takes place at time  $t - \tau_i$ . The branching number  $n_b$  controls the amount of self-excitation and is equal to the average number of events directly triggered by event  $i$ . We also assume  $\langle \mu(t) \rangle < \infty$ , thus the background process dies out eventually.

The dynamics of this process can be equivalently described as a branching processes. Let the root be created at time  $t = 0$  and start generating offsprings with a rate  $\mu(t)$ . Each offspring  $i$ , generated at time  $\tau_i$ , starts generating its own offsprings, with a rate  $n_b \phi(t - \tau_i)$  (see Figure 5). The resulting tree is a Galton-Watson tree with special root offspring distribution. The tree is almost surely finite if  $n_b \leq 1$  [24, 5].

Based on our empirical observations, we specify the functional forms of the kernels as follows:  $\mu(t)$  is given by a Weibull pdf  $\text{Weib}(a, b, \alpha)$  and the memory kernel  $\phi(t)$  is chosen to take the form of a lognormal distribution  $\text{LogN}(\mu, \sigma)$ . The parameters of the model are estimated for each discussion tree separately using a maximum likelihood estimation. Since data contains information on the structure of discussions, we can also separately estimate parameters for the root process and the comments. The general formula for the maximum likelihood function of a non-homogeneous Poisson process with intensity  $\lambda(t)$  is given by

$$\log \mathcal{L}(\tau_1, \tau_2, \dots, \tau_k | \theta) = - \int_0^{\tau_k} \lambda(t) dt + \int_0^{\tau_k} \log(\lambda(t)) dN(t),$$

where  $\tau_1, \dots, \tau_k$  are event times,  $\theta$  is the set of parameters and  $N(t)$  is the counting process [24, 5]. In our setting the loglikelihood function for the intensity  $\mu(t)$  is

$$\begin{aligned} \log \mathcal{L}(\theta(\mu) | \tau) = & -a(1 - \exp(-(\tau_k/b)^\alpha)) - \sum_{i=1}^k ((\tau_k/b)^\alpha + (\alpha - 1) \log(\tau_i)) + \\ & + k(\log(a) + \log(\alpha) - \alpha \log(b)). \end{aligned} \quad (1)$$

The loglikelihood function for  $\phi(t)$  is given as follows

$$\begin{aligned} \log \mathcal{L}(\theta(\phi)|\tau) = & -\frac{1}{2} \left( 1 + \operatorname{erf} \left( \frac{\log \tau_k - \mu}{\sqrt{2}\sigma} \right) \right) + k \log \left( \frac{1}{\sigma\sqrt{2\pi}} \right) \\ & - \sum_{i=1}^k \left( \frac{(\log \tau_i - \mu)^2}{2\sigma^2} + \log \tau_i \right). \end{aligned} \quad (2)$$

Maximization of  $\log \mathcal{L}(\theta|\tau)$  is performed using the L-BFGS-B algorithm, where parameters are constrained to be positive [34].

In order to estimate the branching number  $n_b$  of a given tree, we collect the forward degrees of all comments across the tree and compute the average value:

$$n_b = \frac{1}{n-1} \sum_{v \neq \text{root}} (d_v - 1) = 1 - \frac{d_{\text{root}}}{n-1},$$

where  $d_v$  is the degree of a node  $v$  and  $d_{\text{root}}$  is the degree of the root,

## 4 Model comparison

In this section, we provide an overview of the baseline models and evaluation metrics for model comparison. We evaluate the performance of our model on four sets of discussion trees in Reddit divided by years: (a) 2008, (b) 2010, (c) 2012 and (d) 2014. From each set, we take a subset of all trees of *small* size between 50 and 200 nodes and take a random selection of 8000 trees in each subset. A similar procedure is performed for subsets of *larger* trees of size between 200 and 2000. The following division is used to show similarity of statistical properties of small and large trees.

### 4.1 Modelling structure

To our knowledge, the only model proposed to model the structure of discussion trees is the so-called *preferential attachment (PA) growth model* [9]. Note that this model does not provide information about the temporal evolution of the trees. In this model, trees grow with preferential attachment, as newly arriving nodes connect with a higher probability to the root or to a large degree node. The thread is thus viewed as a growing tree  $\mathcal{T}_n$  with  $n$  nodes  $\{1, 2, \dots, n\}$ , where each node is attached to an existing node  $k$  with probability

$$\mathbb{P}(\text{attach to node } k \mid \mathcal{T}_n) = \frac{1}{Z_n} (\beta_k d_{k,n})^{\gamma_k},$$

and where  $d_{k,n}$  is the degree of node  $k$  in  $\mathcal{T}_n$ ,  $(\beta_k, \gamma_k) = (\beta, \gamma_c)$ , if  $k$  is the root and  $(\beta_k, \gamma_k) = (1, \gamma)$  otherwise, and  $Z_n$  is a normalizing factor. Parameters  $\beta, \gamma_c, \gamma$  govern the strength of preferential



attachment and the root bias, which are fitted for a given tree by maximizing the likelihood of arrival of each new node.

We compare the quality of our model with respect to PA by performing Monte Carlo simulations. For each sample tree we estimate the parameters of the Hawkes model and the PA model and then simulate 50 different trees from each of these models. The PA model takes the number of nodes as an input, thus we let each simulated PA tree have the same number of nodes as the Hawkes tree from the same simulation run to get a proper comparison. A similar strategy was used in [30]. The difference in structure between the given tree  $\mathcal{T}$  and a generated one  $\hat{\mathcal{T}}$  is compared using the mean error per distance layer

$$\varepsilon_d^{\min}(\hat{\mathcal{T}}, \mathcal{T}) = \frac{1}{d_{\min}} \sum_{k=1}^{d_{\min}} |\hat{N}_k - N_k|, \quad \varepsilon_d^{\max}(\hat{\mathcal{T}}, \mathcal{T}) = \frac{1}{d_{\max}} \sum_{k=1}^{d_{\max}} |\hat{N}_k - N_k|,$$

where  $N_k(\hat{N}_k)$  is the number of nodes at distance  $k$  from the root in  $\mathcal{T}(\hat{\mathcal{T}})$  and  $d_{\min}(d_{\max})$  is the minimal (maximal) depth of the trees. Simulation averages of  $\varepsilon_d^{\min}$  and  $\varepsilon_d^{\max}$  are further evaluated for each sample tree.

## 4.2 Predicting activity

Due to the lack of an existing model for the dynamical behaviour of discussions, we consider two models previously designed for Twitter cascades and citation activity with special choice of activity kernels. The evaluation is performed as follows: after fitting the model parameters by observing the timing of events in a learning time window, the task is to predict its future evolution. Parameters evaluation depends on the observation time of the tree evolution. The tree is observed from the appearance of the root (post) at time  $t = 0$  up to the time  $t_{\text{learn}}$  and the parameters are estimated from that truncated data. In our study we use four different learning periods  $t_{\text{learn}}$ : 4, 6, 8 or 12 hours. The data shows that after 12 hours there remains on average less than 20% of activity to predict, thus longer observation windows include almost all the discussion (see insets to Figures 7 and 8, (d)). After learning the parameters, the fitness of the total future temporal activity time series is evaluated through its loglikelihood score of a particular model. We also use Monte-Carlo simulations to determine the average predicted size of the discussion. As above, after 50 runs of simulation we calculate the average predicted size  $\langle \hat{s} \rangle$  and compute the relative size error

$$\varepsilon_s^{t_{\text{learn}}} = \frac{\langle \hat{s} \rangle - s}{s},$$

where  $s$  is the size of a sample tree and  $\hat{s}$  is the predicted size. The following models were used for comparison.

*Dynamic Poisson model (DP)* [4, 2]. The time series is modelled as a point process with deterministic time-dependent intensity  $\lambda(t)$ . DP has originally been applied to Twitter cascades,

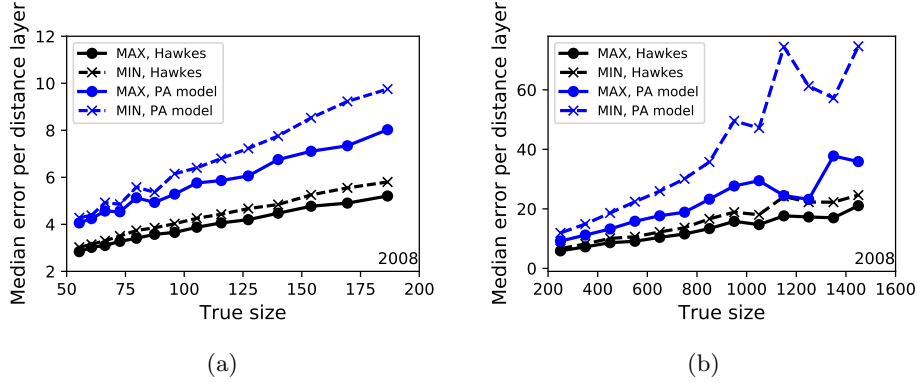


Figure 6: Median profile errors of the node distribution per distance layer for a sample of trees in 2008. The results for small and large trees are depicted in (a) and (b) respectively. MAX error corresponds to  $\varepsilon_d^{\max}$  and MIN stands for  $\varepsilon_d^{\min}$ .

with a power-law functional form for the intensity. In our experiments on discussion trees, we found that the most accurate results were achieved with a lognormal intensity  $\text{LogN}(\mu, \sigma)$ . The parameters are fitted using maximum likelihood method.

*Reinforced Poisson Process (RPP)* [26, 8]. The time series is generated by a point process with reinforced intensity  $\lambda(t, k) = cf(t)r(k)$ , where  $t$  is time and  $k = k(t)$  is the current number of events in the time series. The background intensity  $f(t)$  models the general interest in the subject and was assumed to have a lognormal  $\text{LogN}(\mu, \sigma)$  functional form.  $r(k) = \sum_{i=1}^k \exp(-di)$ , where  $d > 0$ , is the reinforcement factor.

We note the existence of more sophisticated models that use multiplicative Hawkes processes for prediction of Twitter cascades [33, 15]. However, the model [33] is only designed to predict the total size of the cascade, and [15] is designed to model circadian patterns, which are not present in our data.

## 5 Results

The results for the structural modelling are shown in Figure 6. Due to high computational complexity of the inference method of the PA model, we show it only on the example of year 2008. To present the results we divide the trees by size into logarithmically sized bins to obtain a more uniform number of samples in each bin. The median error is computed for each of the bins. The plot shows that both  $\varepsilon_d^{\min}$  and  $\varepsilon_d^{\max}$  grow with the tree size and that the Hawkes model outperforms the PA model. Although for small trees the results are comparable (see Figure 6, (a)), for large trees the gap between  $\varepsilon_d^{\max}$  errors tend to grow drastically (see Figure 6, (b)).

The results for the temporal prediction are illustrated for 2014 (see Appendix for other years). Again, we divide the trees into logarithmically sized bins to obtain a more uniform number of samples in each bin, and calculate the median negative loglikelihood score in each bin. We

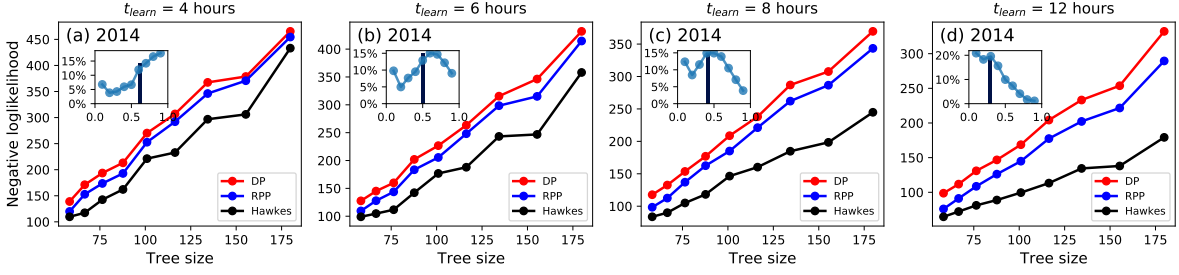


Figure 7: Negative likelihoods for the predicted activity of small trees in 2014. The learning time  $t_{learn}$  for the parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.

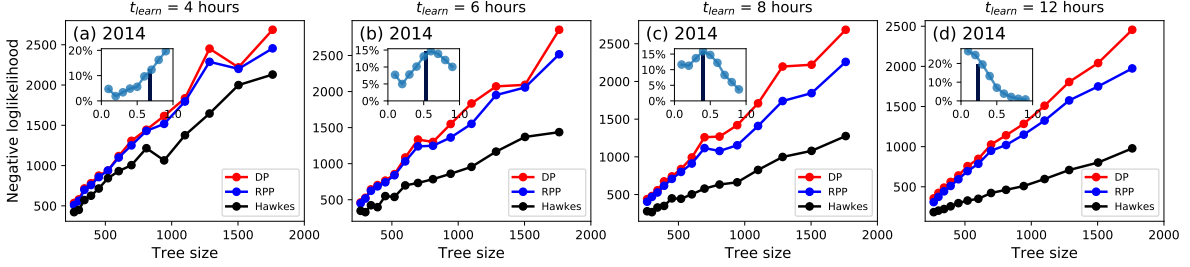


Figure 8: Negative likelihoods for the predicted activity of large trees in 2014. The learning time  $t_{learn}$  for parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.

consistently observe a better fit between the remaining time series and the Hawkes model in the case of small trees, for all considered observation windows  $t_{learn}$  (see Figure 7). In the insets, we observe that for  $t_{learn} = 4$  hours, only 35% of total activity is used for the training. The Hawkes model already shows better fit of the data, however due to the lack of sufficient information, all three models show rather similar performance scores. Longer observation windows result in better parameters estimation and the Hawkes model takes the unbeatable lead for learning times larger than  $t_{learn} = 8$  hours when, on average, around 30% of the cascade is yet to be predicted.

For large trees, a sufficient performance increase is achieved already for  $t_{learn} = 6$  hours (see Figure 8). We may see from the insets that in both cases the methods have to predict similar fraction of activity. Therefore we can state that active commenting period is similar for any post regardless of its size. Larger discussions are thus associated a larger density of events within the discussions life. Since the LV coefficient comes closer to 1, on average, one could hypothesize that bursts and self-excitation become less apparent, and the temporal process approaches a Poisson process. In general, assigning a specific process to the root is an essential ingredient for the prediction.

We show the relative error of total cascade size for samples of small trees in Figure 9. For each observation window  $t_{learn}$  we sample average relative errors and plot a median of these

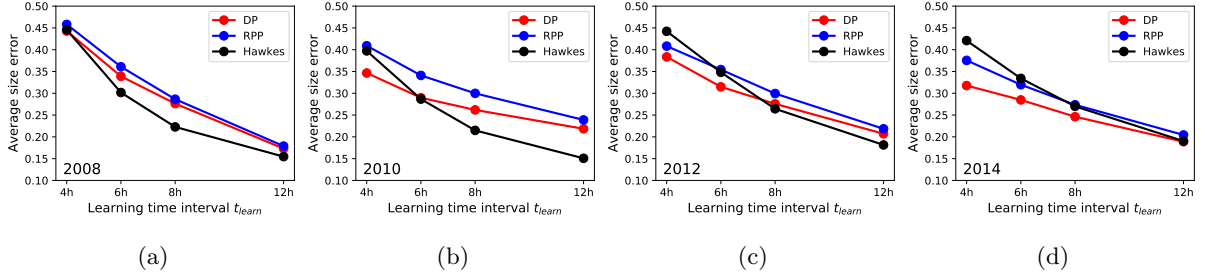


Figure 9: Median relative size errors  $\varepsilon_s^{t_{learn}}$  for the samples of small trees in years (a) 2008, (b) 2010, (c) 2012 and (d) 2014.

values. The short observation window produces larger size prediction error on average, however when  $t_{learn}$  increases, the median error for all models steadily decreases. Although the Hawkes model shows comparable results in total size prediction, it is less accurate for small observation windows, when the lack of information complicates the estimation of a relatively large number of parameters. This limitation could be overcome by a hard setting of some of the variables, thereby reducing dimensionality, or with more sophisticated estimation algorithms.

## 6 Discussion

Online boards play an important role for the exchange of ideas and the occurrence of debates in the online world. In this work, we have proposed a model allowing to reproduce and to predict the structural and temporal properties of discussion trees as they are observed in Reddit. It is important to note here that discussion trees are known to be system-specific and to depend on the user interface and system design of the online board. For example, reply trees observed in Twitter [21] have significantly different properties than those observed in this work. Similarly, changing how information is presented to users visiting the website, e.g. most recent versus most popular, with or without a recommender system based on the user's interests, etc., is expected to result in different tree structures. Nonetheless, the generality of our model, and the fact that Hawkes-type processes have been successfully applied in a variety of social media, suggests that our approach is robust.

As we have discussed, the main advantage of our model is the possibility to predict together, and with a good accuracy, the temporal and structural properties of discussion trees. For each dimension independently, however, some models might provide a better accuracy. For instance, for the modelling of temporal properties, our results have shown that small discussions are better fitted by the Hawkes model than by an ordinary Poisson process. It is evident from Figure 4 that the LV scores for small trees exhibit a larger deviation from Poisson process. However, we also observed that for both small and large trees, the average duration of a post activity is comparable. In large trees, the overall presence of more branches leads to more condensed and more random time series, which may explain the decrease in accuracy of the Hawkes model.

This work opens different venues for future research. For instance, one could improve the prediction power of the model by incorporating the content of posts and comments. It may specifically help to determine the current mood of the discussion and make a conclusion on its future direction [32]. Another way to measure the mood or the popularity of a post is by using its rating. As each post and comment can be rated with ‘likes’ and ‘dislikes’, one could model the dynamics of conflicting debates but also the rating of a post, defined by its number of ‘likes’ minus its number of ‘dislikes’, as a predictor for its future popularity [19]. Collection of the dataset with timestamped ‘likes’ would also provide a deeper insight on the dynamics of opinion formation on the topic.

## Acknowledgement

The authors thank Leto Peel and Daniele Cassese for fruitful discussions and useful suggestions. This work was supported by IAP (Belgian Scientific Policy Office); ARC (Federation Wallonia-Brussels); and by the Russian Foundation of Basic Research [grant number 16-01-00499].

## References

- [1] ADAR, E., AND ADAMIC, L. Tracking information epidemics in blogspace. In *2005 IEEE/WIC/ACM International Conference on Web Intelligence* (2005), pp. 207–214.
- [2] AGARWAL, D., CHEN, B.-C., AND ELANGO, P. Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th International Conference on World Wide Web* (2009), WWW ’09, pp. 21–30.
- [3] CHENG, J., ADAMIC, L., DOW, P. A., KLEINBERG, J. M., AND LESKOVEC, J. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web* (2014), WWW ’14, pp. 925–936.
- [4] CRANE, R., AND SORNETTE, D. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences* 105, 41 (2008), 15649–15653.
- [5] DALEY, D. J., AND VERE-JONES, D. *An Introduction to the Theory of Point Processes Volume I: Elementary Theory and Methods*, second ed. Springer, 2003.
- [6] DOW, P. A., ADAMIC, L. A., AND FRIGGERI, A. The anatomy of large facebook cascades. In *ICWSM’ 13* (2013), pp. 145–154.
- [7] FELLER, W. *An Introduction to Probability Theory and Its Applications*. Wiley, 1968.

- [8] GAO, S., MA, J., AND CHEN, Z. Modeling and predicting retweeting dynamics on microblogging platforms. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (2015), WSDM '15, pp. 107–116.
- [9] GÓMEZ, V., KAPPEN, H. J., AND KALTENBRUNNER, A. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22Nd ACM Conference on Hypertext and Hypermedia* (2011), pp. 181–190.
- [10] GÓMEZ, V., KAPPEN, H. J., LITVAK, N., AND KALTENBRUNNER, A. A likelihood-based framework for the analysis of discussion threads. *World Wide Web* 16, 5-6 (2013), 645–675.
- [11] GRUHL, D., GUHA, R., LIBEN-NOWELL, D., AND TOMKINS, A. Information diffusion through blogspace. In *WWW' 04* (2004), pp. 491–501.
- [12] HAWKES, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58, 1 (1971), 83–90.
- [13] HOFSTAD, R. V. D. *Random Graphs and Complex Networks*, vol. 1 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2016.
- [14] IRIBARREN, J. L., AND MORO, E. Branching dynamics of viral information spreading. *Phys. Rev. E* 84 (2011), 046116.
- [15] KOBAYASHI, R., AND LAMBIOTTE, R. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *ICWSM' 2016* (2016), pp. 191–200.
- [16] KOYAMA, S., AND SHINOMOTO, S. Empirical bayes interpretations of random point events. *Journal of Physics A: Mathematical and General* 38, 29 (2005), L531.
- [17] KUMAR, R., MAHDIAN, M., AND MCGLOHON, M. Dynamics of conversations. In *KDD' 10* (2010), pp. 553–562.
- [18] KWAK, H., LEE, C., PARK, H., AND MOON, S. What is twitter, a social network or a news media? In *WWW' 10* (2010), pp. 591–600.
- [19] LESKOVEC, J., HUTTENLOCHER, D., AND KLEINBERG, J. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web* (2010), WWW '10, pp. 641–650.
- [20] MIURA, K., OKADA, M., AND AMARI, S.-I. Estimating spiking irregularities under changing environments. *Neural Computation* 18, 10 (2006), 2359–2386.
- [21] NISHI, R., TAKAGUCHI, T., OKA, K., MAEHARA, T., TOYODA, M., KAWARABAYASHI, K.-I., AND MASUDA, N. Reply trees in twitter: data analysis and branching process models. *Social Network Analysis and Mining* 6, 1 (2016), 1–13.

- [22] OMI, T., AND SHINOMOTO, S. Optimizing time histograms for non-poissonian spike trains. *Neural Computation* 23, 12 (2011), 3125–3144.
- [23] REDDIT. Official community of /redditmobile, operation start date - march, 31, 2010. <https://www.reddit.com/r/redditmobile/>, Query: 2017-07-13.
- [24] ROSS, S. M. *Stochastic processes*, second ed. Wiley, New York, 1996.
- [25] SANLI, C., AND LAMBIOTTE, R. Local variation of hashtag spike trains and popularity in twitter. *PLOS ONE* 10, 7 (2015), 1–18.
- [26] SHEN, H., WANG, D., SONG, C., AND BARABÁSI, A.-L. Modeling and predicting popularity dynamics via reinforced poisson processes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014), pp. 291–297.
- [27] SHINOMOTO, S., SHIMA, K., AND TANJI, J. Differences in spiking patterns among cortical neurons. *Neural Computation* 15, 12 (2003), 2823–2842.
- [28] STUCK\_IN\_THE\_MATRIX. Dataset is available on the following webpage. <https://files.pushshift.io/reddit/>, Query: 2017-06-01.
- [29] STUCK\_IN\_THE\_MATRIX. I have every publicly available reddit comment for research. approx. 1.7 billion comments @ 250 gb compressed. any interest in this? <https://www.reddit.com/r/datasets/comments/3bxlg7/>, Query: 2017-07-14.
- [30] WANG, C., YE, M., AND HUBERMAN, B. A. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2012), KDD '12, pp. 244–252.
- [31] WANG, D., SONG, C., AND BARABÁSI, A.-L. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.
- [32] YASSERI, T., SUMI, R., RUNG, A., KORNAI, A., AND KERTÉSZ, J. Dynamics of conflicts in wikipedia. *PLOS ONE* 7, 6 (2012), 1–12.
- [33] ZHAO, Q., ERDOGDU, M. A., HE, H. Y., RAJARAMAN, A., AND LESKOVEC, J. 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* (2015), pp. 1513–1522.
- [34] ZHU, C., BYRD, R. H., LU, P., AND NOCEDAL, J. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* 23, 4 (1997), 550–560.

## 7 Appendix

### 7.1 Activity prediction: small trees

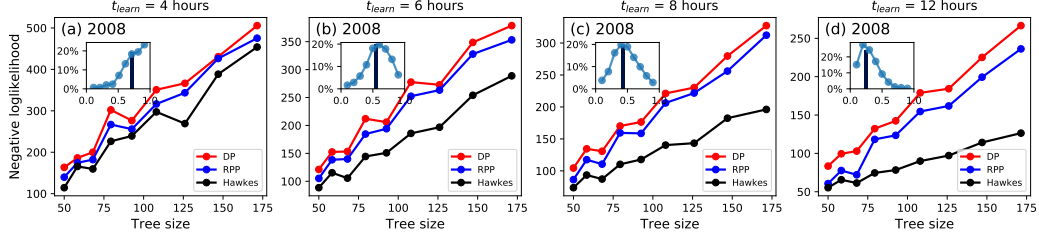


Figure 10: Negative likelihoods for the predicted activity of small trees in 2008. The learning time  $t_{learn}$  for the parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.

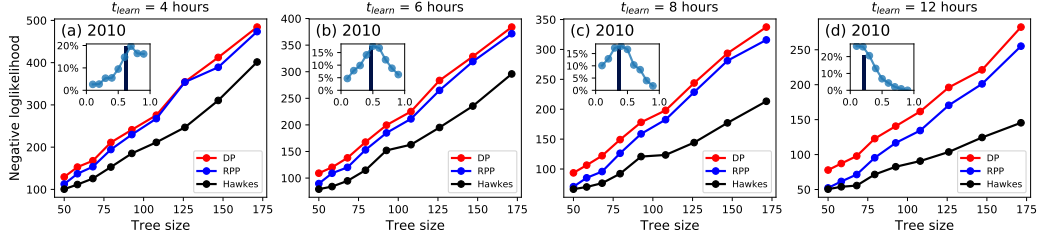


Figure 11: Negative likelihoods for the predicted activity of small trees in 2010. The learning time  $t_{learn}$  for the parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.

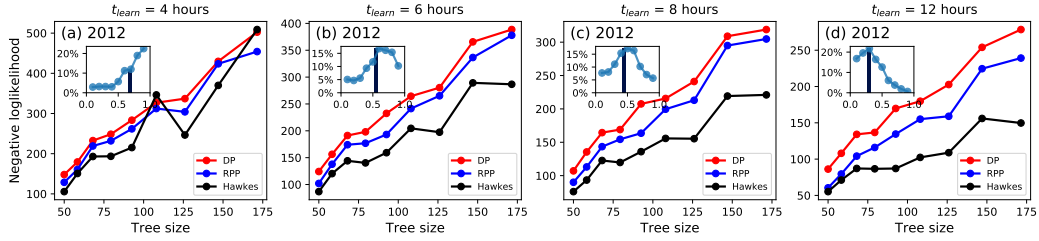


Figure 12: Negative likelihoods for the predicted activity of small trees in 2012. The learning time  $t_{learn}$  for the parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.



## 7.2 Activity prediction: large trees

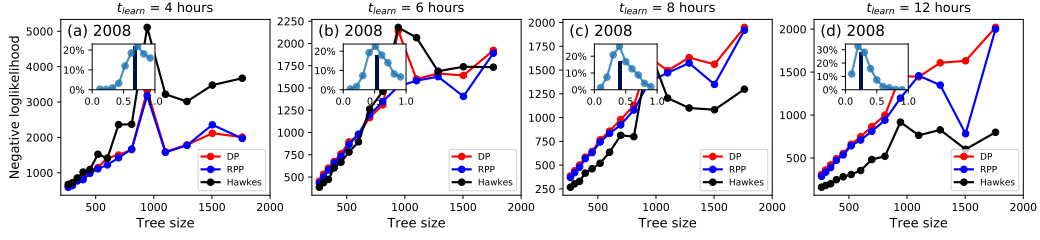


Figure 13: Negative likelihoods for the predicted activity of large trees in 2008. The learning time  $t_{learn}$  for parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.

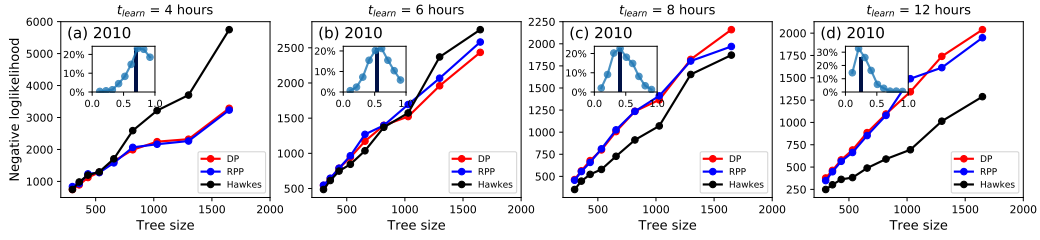


Figure 14: Negative likelihoods for the predicted activity of large trees in 2010. The learning time  $t_{learn}$  for parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.

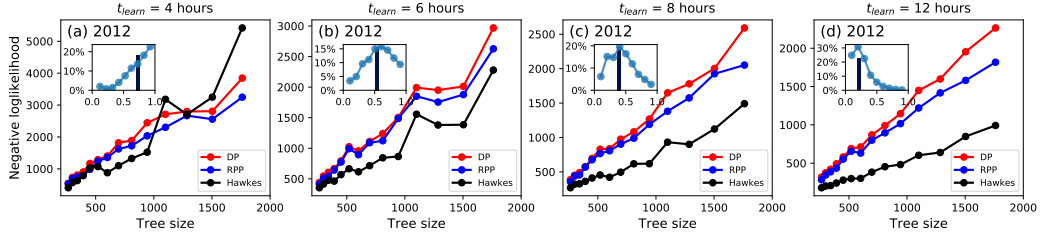


Figure 15: Negative likelihoods for the predicted activity of large trees in 2012. The learning time  $t_{learn}$  for parameters evaluation is (a) 4, (b) 6, (c) 8 and (d) 12 hours. The insets show the distribution of the fraction of activity that remains to predict after the learning phase.