Quantifying uncertainty in a predictive model for popularity dynamics

Joseph D. O'Brien,[1] Alberto Aleta,[2,3] Yamir Moreno,[2,3,4] and James P. Gleeson[1]
[1]*MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick V94 T9PX, Ireland*
[2]*Institute for Biocomputation and Physics of Complex Systems, University of Zaragoza, Zaragoza 50018, Spain*
[3]*ISI Foundation, 10126 Turin, Italy*
[4]*Department of Theoretical Physics, Faculty of Sciences, University of Zaragoza, Zaragoza 50009, Spain*

(Received 26 January 2020; accepted 3 June 2020; published 22 June 2020)

The Hawkes process has garnered attention in recent years for its suitability to describe the behavior of online information cascades. Here we present a fully tractable approach to analytically describe the distribution of the number of events in a Hawkes process, which, in contrast to purely empirical studies or simulation-based models, enables the effect of process parameters on cascade dynamics to be analyzed. We show that the presented theory also allows predictions regarding the future distribution of events after a given number of events have been observed during a time window. Our results are derived through a differential-equation approach to attain the governing equations of a general branching process. We confirm our theoretical findings through extensive simulations of such processes. This work provides the basis for more complete analyses of the self-exciting processes that govern the spreading of information through many communication platforms, including the potential to predict cascade dynamics within confidence limits.

DOI: 10.1103/PhysRevE.101.062311

I. INTRODUCTION

The ease with which individuals may now access online content has revolutionized the way in which information is consumed [1], with social media platforms and online opinion boards [2–5] being two paradigmatic examples of how information is generated, transmitted, and finally absorbed. Modern information communication allow users to both become informed on topics and easily interact online with each other, which may lead to further discussion in a cascadelike manner. This results in the occurrence of tree-shaped cascades, the size of which describes the popularity of a discussion. This size metric is one of the key elements in the modern-day social media sphere and can range from a single post in common discussions to values covering several orders of magnitude for viral events, in an extremely heterogeneous fashion [6–8].

One aspect of modeling information cascades that is of paramount importance, not only on opinion boards but on any type of online media, is the predictability of how content popularity changes over time [9–14]. The ability to accurately forecast which online content will become popular has attracted a great deal of attention for several reasons. First, the sheer volume of content that is now available through online platforms, all of which are competing for the limited attention of the users [15–18], has created a demand for techniques that can decide the order in which users observe new posts, something that is affected by the popularity of the information and at the same time affects that same popularity. This is sometimes referred to as algorithmic bias, by virtue of which more popular topics get more exposure and their popularity is further reinforced. Second, the capacity to predetermine whether or not content will go viral would be an extremely advantageous proposition for corporations, who spend enormous amounts of capital on such campaigns in the hope of increasing their market share [19–21].

Recent literature has focused on developing methods based upon the theory of self-exciting point processes [22]. Unlike homogeneous Poisson processes, the occurrence of previous events in such processes increases the future rate of activity, which can produce a "snowball effect," leading to heavy-tailed distributions of cascade sizes, as observed in empirical data [23–25]. An example of such a process is the Hawkes process [26], the application of which has been used to both model and predict the evolution of retweet cascades on Twitter by incorporating the underlying network topology along with machine learning techniques [27,28], the popularity of threads on Reddit [25,29], and predicting the number of views Youtube videos receive as a result of the discussion taking place on other social media platforms [30]. While these works all highlight the capability of the Hawkes process in predicting popularity cascades of online content, there is a lack of knowledge regarding how the predictability of these cascades is affected by the underlying parameters of the Hawkes process.

The aim of this article is to first introduce an analytically tractable approach to fully describe the Hawkes process using the theory of branching processes, which has been successfully applied to other spreading processes on online media platforms [15–17,31,32]. We obtain the expressions governing the process through an alternative differential-equation framework. Using the results obtained from this analysis, we then turn our attention to the question of predictability in systems with dynamics of this type. We show that our probabilistic interpretation enables the calculation not only of analytical predictions of future popularity but also of levels of confidence in these values directly from the model itself. It should be noted that while we interpret the quantity of

 ©2020 American Physical Society

interest in this work to be the popularity of online content, our general analytic derivation is not restricted to these types of systems. More classically, Hawkes processes have been used to describe a wide range of systems, such as those found in finance [33–38], neuronal activity [39,40], or seismology [41–44]. Hence, the expressions obtained in the following can be applied to any process described by such dynamics simply by changing the phrase discussion tree for point process and post for event.

Before proceeding, we briefly review some of the existing literature that has also considered the use of branching processes in describing such models. The description of an opinion board, specifically Reddit, and also the question of predictability were considered in Ref. [25], wherein parameters estimated from the thread's history were used to calculate the expected number of events in future thread activity. Multiple works have also incorporated other elements of online platforms into their models to assist in predicting cascade sizes. For example, Refs. [14,45] describe Twitter cascades, via extensive numerical predictions, using the follower network of the tweeting individual. In Ref. [46] the authors consider a specific Hawkes process to obtain, through a branching process description, the distribution of events in a certain time window and the distribution of recurrence times between events, without commenting on the predictability of such a model after some observation period.

Our work differs from the aforementioned literature in several ways. First, we consider the most generic form of Hawkes process (unlike, for example, Ref. [45], which makes use of a given memory kernel, or [46], which considers the constant background intensity case) to obtain equations that fully describe the distribution of events at any time; the generality of this approach lends itself to application in a wider number of areas. Second, when we turn to the question of predictability of such a process, which was not considered in [46], we introduce a mathematically tractable model which allows one to calculate the expected size of the underlying tree for any given process while also opening up the possibility of determining prediction intervals for the number of events to occur by a given time. This model thus allows the user to provide a level of confidence in their predictions directly from theory rather than via simulation of an extensive number of ensembles.

The remainder of this paper is organized as follows. In Sec. II we introduce a branching process description of the Hawkes process derived through a differential equation formulation, which is then used to determine properties of cascades generated by such processes. Specifically, our approach allows us to fully determine the distribution of the number of events in a given process at any point in time. Having described the general Hawkes process, in Sec. III we consider the predictability of such processes after observing a cascade over a certain time period. We focus on the case whereby we have observed a given process for some period of time, which we call the observation window, and then, based on the occurrence of events within this window, our theory enables the prediction of not only the expected number of events by some future time, but also the entire distribution of events at each point in the future. In Sec. IV we perform extensive numerical simulations with the aim of validating our model's prediction. We present a summary and our conclusions in Sec. V.

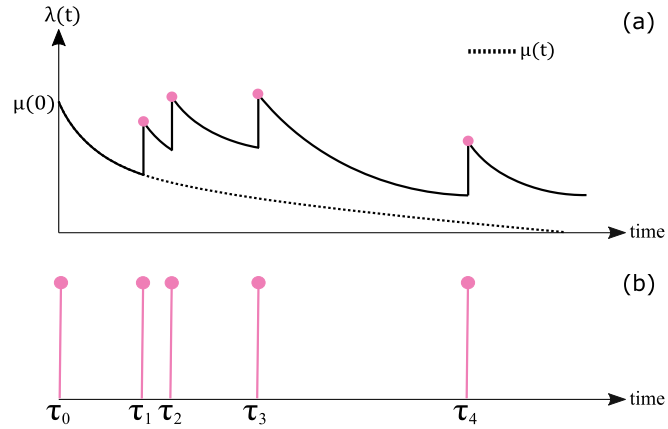

FIG. 1. Example of the Hawkes process. (a) Intensity function $\lambda(t)$ given by Eq. (1) along with the background intensity $\mu(t)$ without any self-excitation, shown by the dashed line. (b) Corresponding event sequence of five events at times τ_i.

II. HAWKES PROCESS MODEL

Self-exciting processes are those in which the likelihood of a stochastic event taking place increases with the occurrence of past events. Hence, the probability of an event occurring in these processes is determined by a time-dependent intensity $\lambda(t)$, defined such that $\lambda(t)dt$ is the expected number of events in the interval $[t, t+dt]$. One particular example of these processes is the Hawkes process [26]. In this class of point processes the intensity is given by

$$\lambda(t) = \mu(t) + \xi \sum_{\tau_i < t} \phi(t - \tau_i), \tag{1}$$

where $\mu(t)$ is known as the background intensity, describing the likelihood of an exogenous event occurring independently of other events. The non-negative $\phi(t)$ term is known as the memory kernel or the excitation function, as it gives the amount by which previous events that have occurred at times $\tau_i < t$ increase the probability of occurrence of future events. As such, it accounts for the endogenous or self-exciting part of the process. Further, we normalize $\phi(t)$ such that $\int_0^\infty \phi(t)dt = 1$. Note that with this interpretation, one may consider $\phi(t)$ as a probability density function or a memory-time distribution [16]. Finally, the parameter ξ may be thought of as the branching number of the process, whereby it defines the average number of events caused as a result of an event and may in some sense capture the fitness of a previous event. Note that in the case $\xi = 0$, one recovers an inhomogeneous Poisson process. In this article we specifically focus on the subcritical case where $\xi < 1$, which results in finite-size (albeit potentially very large) cascades. In the alternative supercritical case ($\xi > 1$) the average number of events grows exponentially without bound, which is evidently meaningless for the processes of interest in this work. Figure 1 shows an example realization of a Hawkes process.

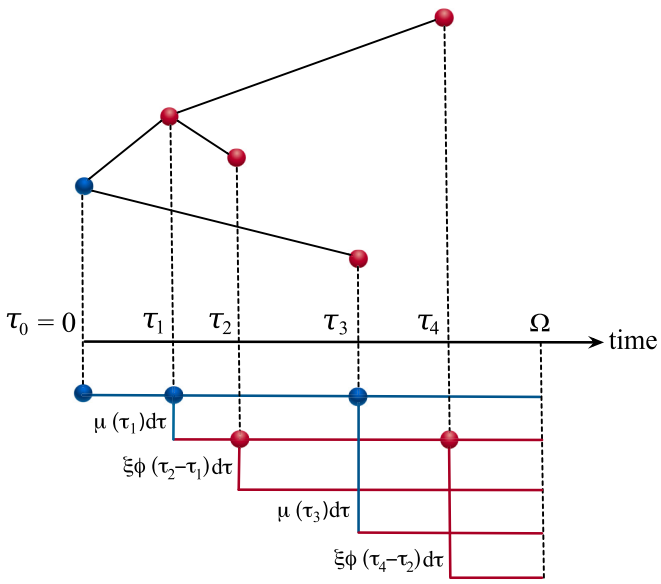

FIG. 2. Schematic of the model. A discussion is started by the seed at time $\tau_0 = 0$, represented by the horizontal blue line. At time τ_1 the seed receives a reply, shown by the blue circle, with probability $\mu(\tau_1)d\tau$, which starts its own subtree shown by the top horizontal red line. At time τ_2 the comment receives a reply highlighted by the red circle, with probability $\xi\phi(\tau_2 - \tau_1)d\tau$. The seed receives another reply at time τ_3, with probability $\mu(\tau_3)d\tau$, which again starts another subtree. Finally, the subtree started at time τ_1 receives another reply at time τ_4, which occurred with probability $\xi\phi(\tau_4 - \tau_1)d\tau$, itself starting a new branch. We now observe the entire tree size, i.e., the number of comments, to be 5 at time Ω.

To describe this process further we will proceed to consider the behavior of a discussion tree on a social media platform. Suppose that a discussion is started by a user posting at time $\tau_0 = 0$. This comment will be referred to as the seed of the discussion. The size of the discussion can increase in two ways, as described in [25]. First, it may receive new comments itself, which are essentially replies to the seed, with rate $\mu(t)$. This rate can depend on a number of factors, such as the discussion's position on the "front page" of the site [47,48] or the number of users present on the site in the period after the tree was created. Second, comments made to the seed can themselves receive replies and thus be viewed as subtrees of the tree that represents the whole discussion. A comment made at time τ_i receives replies with intensity $\xi\phi(t - \tau_i)$, where ξ can be viewed as some fitness parameter which could depend on factors such as the quality of the content in the comment, the time of day at which it first appeared, or the popularity of the individual who made the comment. For simplicity, however, we will consider it as a parameter unique to each discussion.

We can graphically arrange the posts in a discussion by the shape of a tree in which the nodes represent posts and an edge is established between two nodes if one is a reply to the other (see Fig. 2). This representation suggests that the dynamics of such a system can be described as a branching process. Letting $q_m(t)$ be the probability that a discussion of age t has popularity m, i.e., it has received m comments, we may define the probability generating function (PGF) [49] of the popularity distribution as

$$H(t;x) = \sum_{m=1}^{\infty} q_m(t)x^m. \tag{2}$$

In passing we mention one important property of PGFs that we will frequently use in the following, namely, that the PGF for the sum of two random variables X and Y is simply the product of their individual generating functions. For the remainder of this section, we will focus on a derivation of this PGF through a branching process description of the Hawkes process [50,51], followed by an analysis of some of the properties that can be extracted from this interpretation.

A. Branching process description of the Hawkes process

We propose a derivation of the equations for the branching process generating function given by Eq. (2) that is based on considering the change in the PGFs over infinitesimal time intervals. This leads naturally to a differential equation formulation. In the Appendix we show how the method introduced below can also be used to derive the well-known integral equation for the Bellman-Harris process [31,32,50,51].

Let us first consider the distribution of the size of subtrees described above. Suppose that a new comment arrives at the subtree at time τ. We define the PGF of the resulting size distribution as $G(\tau, a, \Omega; x)$, where a is the age of the comment that originated the subtree and Ω the time at which the distribution is observed. Note that the time of origin of the subtree is thus $c \equiv \tau - a$. Now suppose we examine how this PGF changes over a time interval $(\tau - \Delta t, \tau)$. If Δt is sufficiently small so that at most one event occurs in a time interval of length Δt, there are two possible events that may occur.

(i) A user may decide to reply to the comment of age a with probability $\xi\phi(a)\Delta t$, which depends on both the time elapsed since the comment was made and its fitness. This will result in a new comment of age 0 which may in turn receive its own replies. Concurrently, the original comment may also spawn further replies, so the total contribution to $G(\tau - \Delta t, a - \Delta t, \Omega; x)$ is $G(\tau, 0, \Omega; x)G(\tau, a, \Omega; x)$, where we have made use of the aforementioned property that the PGF for the sum of two random variables is simply the product of their respective PGFs.

(ii) There may also be no replies to the comment in this time interval. This occurs with probability $1 - \xi\phi(a)\Delta t$, and the contribution to $G(\tau - \Delta t, a - \Delta t, \Omega; x)$ in this case is simply $G(\tau, a, \Omega; x)$.

Thus, the equation governing the change in G over the interval $(\tau - \Delta t, \tau)$ is given by

$$G(\tau-\Delta t, a-\Delta t, \Omega; x) = \xi\phi(a)\Delta t G(\tau, 0, \Omega; x)G(\tau, a, \Omega; x) + [1 - \xi\phi(a)\Delta t]G(\tau, a, \Omega; x). \tag{3}$$

This equation may then be expressed as a partial differential equation by considering the two-dimensional Taylor approximation in the small-Δt limit to obtain

$$\frac{\partial G}{\partial \tau} + \frac{\partial G}{\partial a} = \xi\phi(a)[1 - G(\tau, 0, \Omega; x)]G, \tag{4}$$

where we write G for $G(\tau, a, \Omega; x)$. This equation may be simplified via the method of characteristics [52] (letting

$c = \tau - a$ be constant along a characteristic) to obtain the ordinary differential equation

$$\frac{dG}{d\tau} = \xi\phi(\tau - c)[1 - G(\tau, 0, \Omega; x)]G \tag{5}$$

for the change in G along a characteristic, which we may solve by noting that $G(\Omega, \Omega - c, \Omega; x) = x$, i.e., the initial subtree size when a comment is made is 1, to obtain

$$G(\tau, \tau - c, \Omega; x) = x \exp\left\{-\xi \int_{\tau}^{\Omega} \phi(v - c)[1 - G(v, 0, \Omega; x)]dv\right\}. \tag{6}$$

We may now insert $a = \tau - c$ and introduce the "tree age" $t = \Omega - \tau$, i.e., the age of the subtree at observation time, and let $w = \Omega - v$, which gives

$$G(t, a; x) = x \exp\left\{\xi \int_0^t \phi(w + a)[G(t - w, 0; x) - 1]dw\right\}. \tag{7}$$

Finally, we let $a = 0$ to obtain the PGF for the subtree size at age t as a result of a new post (and dropping the $a = 0$ argument from G)

$$G(t; x) = x \exp\left\{\xi \int_0^t \phi(w)[G(t - w, x) - 1]dw\right\}. \tag{8}$$

We will now proceed to determine the PGF of the entire size distribution of a tree with age t, $H(t; x)$. Equation (8) describes the distribution of the size of a subtree as a result of replying to either the original seed or another comment (which would be itself a root of a previous subtree). To determine the entire tree size distribution we first define $H(\tau, \Omega; x)$ to be the PGF of a cascade that had started at time 0, observed over the time interval (τ, Ω). As above, we consider the two possible events which may occur in relation to the seed over the small time interval $(\tau - \Delta t, \tau)$.

(i) There may be a comment made on the post (i.e., in direct reply to the seed) with probability $\mu(t)\Delta t$, resulting in a subtree that develops from this comment with size distribution determined by $G(\tau, 0, \Omega; x)$. The seed itself may also receive further comments. The total contribution to $H(\tau - \Delta t, \Omega; x)$ is given by the product of the two individual PGFs $G(\tau, 0, \Omega; x)H(\tau, \Omega; x)$.

(ii) The seed may receive no comments in this time interval which occurs with probability $1 - \mu(\tau)\Delta t$, so the contribution to $H(\tau - \Delta t, \Omega; x)$ is simply $H(\tau, \Omega; x)$.

Analogously to the previous PGF, we take the $\Delta t \to 0$ limit and use the method of characteristics to obtain the differential equation

$$\frac{dH(\tau, \Omega; x)}{d\tau} = \mu(\tau)[1 - G(\tau, 0, \Omega; x)]H(\tau, \Omega; x), \tag{9}$$

noting that the initial tree size is 1 (the seed itself), i.e., $H(0, 0; x) = x$. We remark that Ω is in fact the age of the post. For consistency, we will define this to be t and set $\tau = 0$ such that we observe the entire tree, obtaining

$$H(t; x) = x \exp\left\{\int_0^t \mu(y)[G(t - y; x) - 1]dy\right\}, \tag{10}$$

where $G(t - y; x)$ is given by Eq. (8). Therefore, the entire behavior of the Hawkes process is described by Eqs. (8) and (10). Note that a similar expression was obtained by Hawkes in the case of constant background intensity, i.e., $\mu(t) = \lambda_0$, by considering the process as a Poisson cluster process [53]; however, a description of the most general case including time-varying background intensity is lacking.

B. Mathematical analysis and properties of the branching tree

1. Mean size of a cascade

To determine the expected size of a discussion with age t, $m(t)$, we use the property that

$$m(t) = \sum_{n=1}^{\infty} nq_n(t) = \left.\frac{\partial H(t; x)}{\partial x}\right|_{x=1}. \tag{11}$$

We then differentiate Eq. (10) and evaluate at $x = 1$ to obtain an integral equation for $m(t)$,

$$m(t) = 1 + \int_0^t \mu(y)m_G(t - y)dy, \tag{12}$$

where $m_G(t)$ represents the mean size of a subtree with age t and is defined by $m_G = \frac{\partial G}{\partial x}|_{x=1}$. Analogously, differentiating Eq. (8) yields

$$m_G(t) = 1 + \xi \int_0^t \phi(w)m_G(t - w)dw. \tag{13}$$

The next step is to take the Laplace transform of Eq. (12) to obtain

$$\hat{m}(s) = \frac{1}{s} + \hat{\mu}(s)\hat{m}_G(s), \tag{14}$$

with $\hat{m}_G(s)$ given by the Laplace transform of Eq. (13), which may be solved exactly to obtain

$$\hat{m}_G(s) = \frac{1}{s[1 - \xi\hat{\phi}(s)]}; \tag{15}$$

substituting this value in Eq. (14) gives

$$\hat{m}(s) = \frac{1}{s} + \frac{\hat{\mu}(s)}{s[1 - \xi\hat{\phi}(s)]}. \tag{16}$$

We may determine the limiting value of $m(t)$ as $t \to \infty$, if it exists, by making use of the final limit theorem [54]

$$m(\infty) = \lim_{t\to\infty} m(t) = \lim_{s\to 0} s\hat{m}(s) \tag{17}$$

$$= 1 + \frac{\int_0^\infty \mu(t)dt}{1 - \xi}. \tag{18}$$

Furthermore, it is also interesting to understand how this limit is approached in the large-t limit. Following [16,32], the calculation depends on the existence (or not) of the value α that satisfies

$$\xi \int_0^\infty e^{-\alpha t}\phi(t)t = 1. \tag{19}$$

This special value of α is known as the Malthusian parameter [50]. If this parameter exists then the limit $m(\infty)$ is

approached exponentially:

$$m(t) \sim m(\infty) + \frac{\int_0^\infty e^{-\alpha p}\mu(p)dp}{\alpha\xi \int_0^\infty p\,\phi(p)e^{-\alpha p}dp} e^{\alpha t} \quad \text{as } t \to \infty. \tag{20}$$

However, if there is no solution α of Eq. (19), which is the case when $\phi(t)$ is a subexponential distribution (e.g., Gamma distribution), then the large-t behavior of $m(t)$ is given by

$$m(t) \sim m(\infty) - \frac{\xi \int_0^\infty \mu(p)dp}{(1-\xi)^2}[1 - C(t)], \tag{21}$$

where $C(t) = \int_0^t \phi(t)dt$ is the cumulative distribution of the memory kernel.

While the above analysis describes the mean behavior of the long-time dynamics, for some choices of the background intensity and memory kernel it is also possible to exactly determine the entire temporal behavior of the mean cascade size. In the following we will explore two of those choices: constant background intensity with exponential memory kernel and constant background intensity with power-law memory kernel.

a. Exponential memory kernel with constant background intensity. To consider a specific example, we take the well-studied case where the process's memory kernel is exponentially distributed [26,53,55–57] with mean time $1/\beta$ and background intensity given by a constant, i.e., $\mu(t) = \lambda_0$, so that the intensity is given by

$$\lambda(t) = \lambda_0 + \xi \sum_{\tau_i < t} \beta e^{-\beta(t-\tau_i)}; \tag{22}$$

in this case (16) becomes

$$\hat{m}(s) = \frac{1}{s} + \frac{\lambda_0}{s^2}\left[\frac{\beta + s}{s + \beta(1-\xi)}\right], \tag{23}$$

which results in the following solution for the mean cascade size as a function of time:

$$m(t) = 1 + \frac{\lambda_0}{1-\xi}\left[t + \frac{\xi}{\beta(1-\xi)}(e^{-\beta(1-\xi)t} - 1)\right]. \tag{24}$$

Note that we initially have a linear term in t, as one would expect in the case of constant background intensity, which gives some constant probability of replies to a post regardless of the time. This growth does eventually slow down as a result of the exponential decay term, but is still determined as linear growth. A detailed analysis into this form of the Hawkes process was considered in [58], where it was shown that this simplistic setting is equivalent to a continuous-time Markov process, as one would anticipate from the exponential kernel.

b. Shifted power-law memory kernel with constant background intensity. Another frequently used kernel is that of the shifted power-law memory kernel $\phi(t) \propto (t + c)^{-(1+\beta)}$ for $\beta > 0$ [14,30,41,59–61]. From this kernel, along with a constant background intensity and the normalization condition [i.e., $\int \phi(t)dt = 1$] we have

$$\lambda(t) = \lambda_0 + \xi \sum_{\tau_i < t} \beta c^\beta (t - \tau_i + c)^{-(1+\beta)}. \tag{25}$$

Hence, in this case $\hat{\mu}(s) = \frac{\lambda_0}{s}$ and $\hat{\phi}(s) = \beta c^\beta e^{cs} s^\beta \Gamma(-\beta, cs)$, where Γ is the upper incomplete Gamma function. We may determine the early-time behavior in this instance by considering $s \gg 1$ and expanding the incomplete Γ function giving

$$\hat{\phi}(s) \sim \frac{\beta}{c}\frac{1}{s}. \tag{26}$$

Inserting this expression into Eq. (16) and applying the inverse Laplace transform, we obtain

$$m(t) \approx 1 + \frac{c\lambda_0}{\xi\beta}[e^{(\beta/\xi c)t} - 1]. \tag{27}$$

The presence of a constant background intensity makes the question regarding long-time behavior less relevant. This is due to the fact that the constant (nonzero) probability of an event to occur results in infinitely large cascades in the large-t asymptotic limit described, on average, by Eq. (18). If this factor were not constant but rather a time-decaying $\mu(t)$, as the theory allows, the large-t behavior would be governed by Eq. (21).

2. Probability a discussion receives no comments

Another interesting quantity which may be determined directly from the probability generating function is the probability that a discussion tree, once started by a user, receives no responses by the time it has age t, denoted given by $q_1(t)$. This value can be obtained from Eq. (10) as follows:

$$q_1(t) = \lim_{x\to 0} \frac{H(t;x)}{x} \tag{28}$$

$$= \lim_{x\to 0} \exp\left\{\int_0^t \mu(y)[G(t-y;x) - 1]dy\right\} \tag{29}$$

$$= \exp\left\{-\int_0^t \mu(y)dy\right\}. \tag{30}$$

As one would expect, this quantity is purely determined by the background intensity function as the memory kernel only becomes a factor once the seed has received a reply.

3. Distribution of the number of events

The branching process interpretation of a given Hawkes process described in Sec. II A allows one to determine the entire dynamics of the given process through its PGFs (8) and (10). In order to extract $q_m(t)$, that is, the probability that a tree has size m at age t, from these equations, one is required to differentiate the PGF $H(t;x)$ m times, i.e.,

$$q_m(t) = \frac{1}{m!}\frac{\partial^m}{\partial x^m} H(t;x)\bigg|_{x=0}; \tag{31}$$

the numerical differentiation required to perform this calculation is however inaccurate for large-m values and as such we instead use the Cauchy formula [62] for the derivative of a function to obtain

$$q_m(t) = \frac{1}{2\pi i}\oint_C H(t;x)x^{-(m+1)}dx, \tag{32}$$

where C is a contour in the complex-x plane such that all poles of $H(t;x)$ lie outside C. This may then be evaluated in a numerically accurate manner through inverse fast Fourier transform routines [15,62,63] to allow one to obtain the distribution of tree sizes at a given time t.

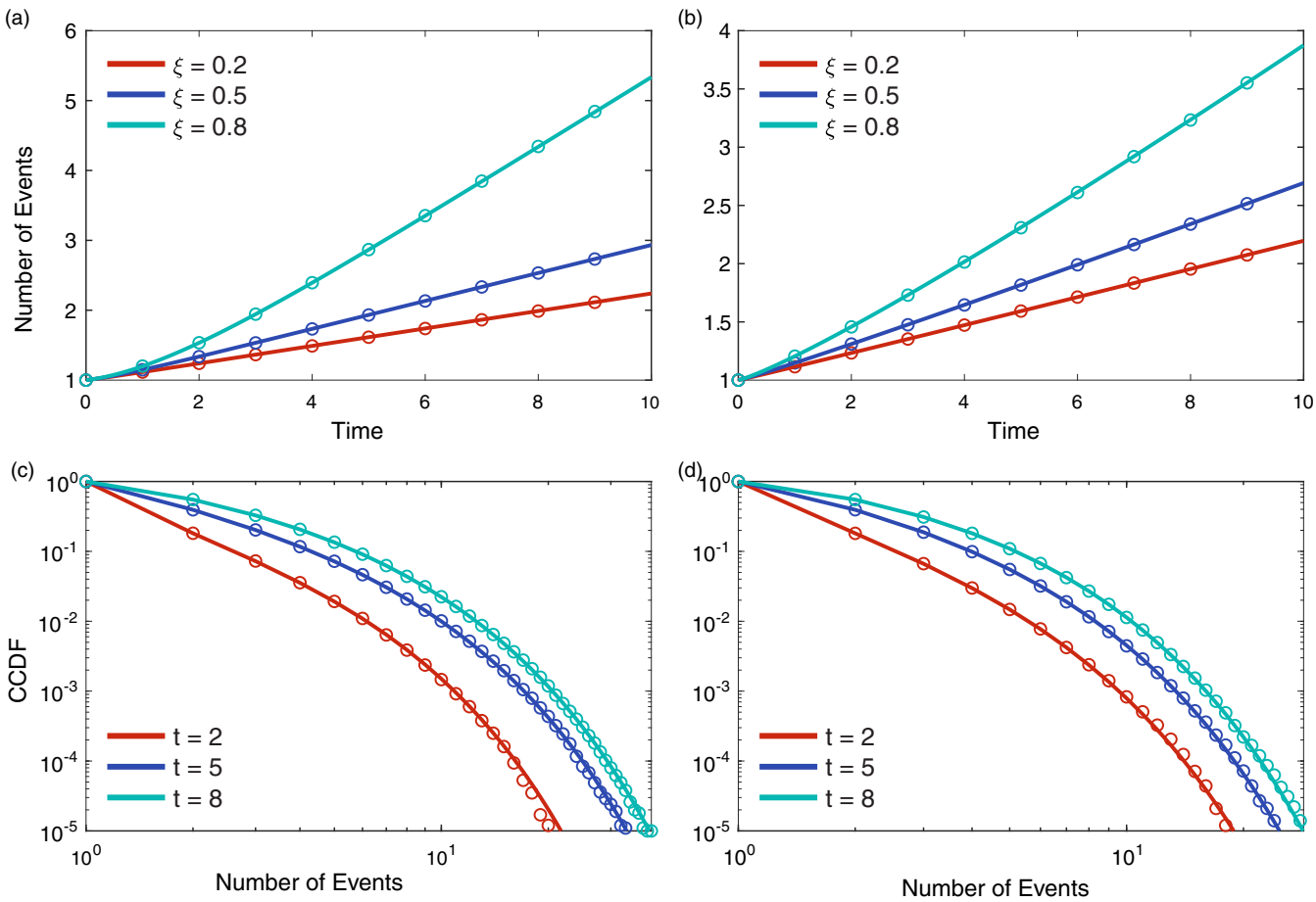

FIG. 3. Numerical simulation of the Hawkes process over 10^6 realizations. (a) Mean number of events for a Hawkes process with exponential memory kernel ($\beta = 3$) and constant background intensity $\lambda_0 = 0.1$ for a range of fitness values. The lines represent the theoretical values for the expected number of events given by Eq. (24), while the circles represent the corresponding results from simulation. (c) The CCDF of the number of events in a scenario where $\xi = 0.5$ and $\beta = 3$ for three different values of observation time. Again, the lines represent the theoretical CCDF from inversion of the PGF and circles represent the results from simulation. (b) and (d) Corresponding plots for a power-law memory kernel with $c = 0.01$ and $\beta = 0.3$.

C. Numerical simulations

In Fig. 3 we compare the theoretical predictions with numerical simulations of the Hawkes process. In Figs. 3(a) and 3(c) we consider the case of constant background intensity and exponential memory kernel. Specifically, in Fig. 3(a) the expected number of events is shown for a number of different fitness parameters ($\xi = 0.2, 0.5, 0.8$) along with $\lambda_0 = 0.1$ and $\beta = 3$. The lines correspond to the theoretical expected value of Eq. (24), while the circles represent the average number of events over an ensemble of 10^6 realizations. As expected, larger fitness values result in a greater expected number of events. Note also that for large values of t, a linear behavior of the expected value is recovered, as suggested by the theoretical result of Eq. (24).

In each case we extract the distribution of tree sizes at each time point via the approach described in Sec. II B 3. Figure 3(c) shows the corresponding complementary cumulative distribution function (CCDF), which is the probability of having m or more comments on the discussion, following the process described above, at several different ages. Regarding these results, it is clear that the likelihood of a larger number of events is increasing with time.

Figures 3(b) and 3(d) show the equivalent tests for the case of the power-law memory kernel given by Eq. (25) with $\beta = 0.3$ and $c = 0.01$. Note that although there is no closed-form expression for the expected value in Fig. 3(b), we can determine these values directly from the distribution function obtained through the PGF of the process. We find the results to be in strong agreement with simulations (in spite of the limitations of discrete-time simulations of such dynamics [64]), therefore validating the branching process interpretation.

One may note that the expected size of cascades shown in Fig. 3 is not converging to a fixed value; this is due to the constant background intensity, which results in some probability of an event occurring at each point that leads to the cascades growing, on average, without bound as suggested by Eq. (18). If instead the background intensity is decaying with time, the possibility of convergence on average to a fixed value is more evident. In Fig. 4 we consider a purely self-exciting process such that $\lambda(t) = \xi\phi(t)$, specifically the same exponential memory kernel as in Fig. 3 with $\xi = 0.8$ is used. Figure 4(a)

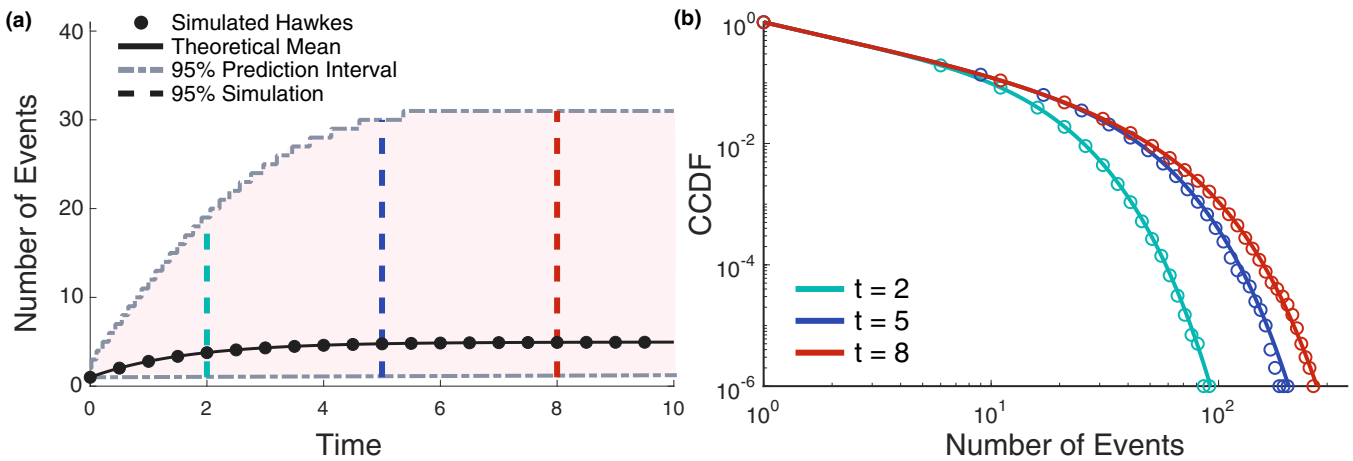

FIG. 4. Numerical simulation of the Hawkes process over 10^6 realizations. (a) Mean number of events for a purely self-exciting Hawkes process with exponential memory kernel ($\beta = 3$) and equivalent background intensity with a fitness value of $\xi = 0.8$. The black solid line represents the theoretical values for the expected number of events, while the circles represent the corresponding results from simulation. The gray dash-dotted lines indicates the 95% prediction interval, i.e., the values between which 95% of the distribution is centered at each time point, while the colored dashed lines indicate the equivalent percentiles from simulation at three time points $t = 2, 5, 8$. (b) The CCDF of the number of events at the same three time points.

shows the expected number of events and also the 95% prediction intervals, i.e., the values within which we would expect 95% of ensembles appeared. The dashed color lines show the equivalent percentiles from simulations. We note that in this scenario these quantities converge at large time; for example, the expected number of events tends towards 5 as Eq. (18) indicates. Figure 4(b) shows the corresponding CCDFs at each of the time points considered in Fig. 4(a). We also comment on the wideness of the prediction intervals describing the process which highlight how they are extremely heterogeneous in comparison to the expected number of events. This highlights the inherent stochasticity underlying processes described by such dynamics, which further emphasizes the importance in obtaining levels of confidence in predictions aside from simply the expected number of events that appears to be an insufficient predictor by itself, which is particularly important when such methods are applied to empirical data.

III. PREDICTABILITY OF DISCUSSION POPULARITY

We now focus on the issue of predicting the time-dependent discussion size given that its popularity has been observed from the moment it was first created at time $\tau_0 = 0$, until a time T, i.e., over the tree age interval $[0, T]$, which we will refer to as the observation window (see Fig. 5). Let us assume that we have observed n comments over this interval at times $\{\tau_0, \tau_1, \ldots, \tau_{n-1}\}$ and that we now wish to consider how this process evolves until a new observation at time Ω, i.e., the tree age interval $(T, \Omega]$, which we refer to as the prediction window. Consistent with the branching process equivalence for the Hawkes process, we may treat each of these subtrees seeded during the observation window as independent and we only consider what occurs in that subtree after the time T. This means that while a comment may have resulted in a new subtree before time T, we treat both of them independently when they develop after time T such that at time T we have n subtrees all of size one but with different ages given by $T - \tau_i$, where τ_i denotes the time when the subtree was seeded.

As in Sec. II A, we have to consider two different types of events. First, we have the seed which started the discussion. Then the remaining $n - 1$ comments are replies to the seed or to previous posts on the discussion. We again start by defining $G(\tau, a, \Omega; x)$ to be the PGF of the size distribution of a subtree, seeded as a result of a reply at time τ to a comment which had age a at this time, observed at time Ω. We are now interested in predicting the dynamics of these $n - 1$ subtrees over the time interval $(T, \Omega]$. To do so, let us consider the $(i + 1)$th comment which was made at time $\tau_i < T$ and study the size of the subtree resulting from a reply to this comment at time τ. Therefore, Eq. (6) in this case is

$$G(\tau, a, \Omega; x) = x \exp\left\{-\xi \int_{\tau}^{\Omega} \phi(v + a - \tau)[1 - G(v, 0, \Omega; x)]dv\right\}. \tag{33}$$

The age of the original comment at this time will be $a = \tau - \tau_i$. As we are interested in considering the tree size over the entire prediction window, we set $\tau = T$ so that

$$G(T, T - \tau_i, \Omega; x) = x \exp\left\{-\xi \int_{T}^{\Omega} \phi(v - \tau_i)[1 - G(v, 0, \Omega; x)]dv\right\}. \tag{34}$$

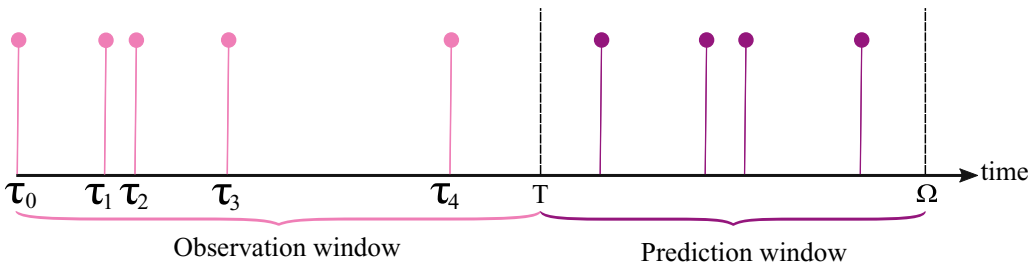

FIG. 5. Schematic of the model described in Sec. III. For a given process, we observe its evolution over what we refer to as the observation window from the time τ_0 of the seed comment until a future time T. During this window we observe $n = 5$ events, including the seed event itself, at times $\{\tau_0, \tau_1, \ldots, \tau_4\}$. Our model then considers how the age of each of these events, $a_i = T - \tau_i$, influences the distribution of number of events over a prediction window of length $r = \Omega - T$, where Ω is the time at which we observe what has occurred in this window. In the ensemble shown above, four further events took place by the observation time.

Next we define the prediction age, i.e., the length of the prediction window, as $r = \Omega - T$. Similarly, letting $w = \Omega - v$, we can rewrite the function in terms of tree age to abridge notation, obtaining

$$G(r, T - \tau_i; x) = x \exp\left\{-\xi \int_0^r \phi(r + T - \tau_i - w)[1 - G(w, 0; x)]dw\right\}. \tag{35}$$

We note that the G function within the integral, i.e., for a comment that was made during the prediction window, is exactly of the form described by Eq. (8). The G function described by the above equation, however, has an explicit dependence on the age of the comment when we stop observing, $a_i = T - \tau_i$. Thus, in order to compress notation, we define $G_i(r; x) = G(r, a_i; x)$ to highlight the difference between trees started during the observation and prediction windows.

We must also consider replies to the seed comment over this interval. A similar line of thought allows one to see that the PGF describing the distribution of these replies is given by

$$H(r; x) = x \exp\left\{-\xi \int_0^r \mu(r + a_0 - w)[1 - G(w, 0; x)]dw\right\}, \tag{36}$$

where a_0 is simply the age of the discussion when we stop observing, i.e., $a_0 = T$.

Finally, we require one more PGF to completely define the evolution of this process. We denote by $I(\{a_0, a_1, \ldots, a_{n-1}\}, r; x)$ the PGF of total discussion tree size after a prediction window of length r, where we have observed n comments on the tree which had ages $\{a_0, a_1, \ldots, a_{n-1}\}$ when we stopped observing (i.e., at $t = T$). Therefore, it can be expressed as

$$I(\{a_0, a_1, \ldots, a_{n-1}\}, r; x) = H(r; x) \prod_{i=1}^{n-1} G_i(r; x) \tag{37}$$

$$= x^n \exp\left[-\int_0^r \left(\mu(r + a_0 - w) + \sum_{i=1}^{n-1} \xi\phi(r + a_i - w)\right)[1 - G(w, 0; x)]dw\right], \tag{38}$$

where the product of generating functions arises as we are summing the contributions from each of the subtrees started before time T. Note that this equation reduces to Eq. (10) in the case where we observe only one event ($n = 1$), which has age 0 when we stop observing ($a_i = 0$) and letting $t = r$.

A. Mean popularity of predicted cascade size

To determine the mean tree size of such a discussion, we can differentiate Eq. (38) and evaluate at $x = 1$ as in Sec. II. This gives

$$m_I(r) = \left.\frac{\partial I(\{a_0, a_1, \ldots, a_{n-1}\}, r; x)}{\partial x}\right|_{x=1} \tag{39}$$

$$= n + \int_0^r \left(\mu(r + a_0 - w) + \sum_{i=1}^{n-1} \xi\phi(r + a_i - w)\right) \times m_G(w)dw, \tag{40}$$

where the n term represents the events that occurred during the observation window and $m_G(w)$ represents the mean tree size of prediction age r and is given by Eq. (13). Observing that

$$\xi \int_0^r \phi(r + a_i - w)m_G(w)dw \tag{41}$$

can be written as a convolution integral in terms of the quantity $\Psi_i(r)$ defined by $\Psi_i(r) = \phi(r + a_i)$, we note that the term in Eq. (41) has Laplace transform $\xi\,\hat{\Psi}_i(s)\,\hat{m}_G(s)$. Applying this to each term in the summation of Eq. (40) gives the Laplace transform of $m_I(r)$ as

$$\hat{m}_I(s) = \frac{n}{s} + \frac{1}{s}\left[\frac{\hat{\Phi}(s) + \sum_{i=1}^{n-1} \xi\,\hat{\Psi}_i(s)}{1 - \xi\hat{\phi}(s)}\right], \tag{42}$$

where

$$\hat{\Psi}_i(s) = \int_0^\infty \phi_i(r + a_i)e^{-rs}dr \tag{43}$$

$$= e^{a_i s}\left[\hat{\phi}(s) - \int_0^{a_i} \phi(\tilde{r}) e^{-\tilde{r}s} d\tilde{r}\right], \tag{44}$$

using the change of variable $\tilde{r} = r + a_i$. Similarly, we define $\Phi(r) = \mu(r + a_0)$ such that

$$\hat{\Phi}(s) = e^{a_0 s}\left[\hat{\mu}(s) - \int_0^{a_0} \mu(\tilde{r}) e^{-\tilde{r}s} d\tilde{r}\right]. \tag{45}$$

The evaluation of this transform now depends only on (a finite) numerical integration, provided we know the Laplace transform of both ϕ and μ. Thus, this approach is extremely feasible. Furthermore, we can also determine some information regarding the large-r behavior of the mean via the final-value theorem to obtain

$$\lim_{r\to\infty} m_I(r) = \lim_{s\to 0} s\hat{m}_I(s) \tag{46}$$

$$= n + \left(\frac{\hat{\Phi}(0) + \xi \sum_{i=1}^{n-1} \hat{\Psi}_i(0)}{1-\xi}\right), \tag{47}$$

where by Eq. (44)

$$\hat{\Psi}_i(0) = 1 - \int_0^{a_i} \phi(\tilde{r}) d\tilde{r} \tag{48}$$

and similarly

$$\hat{\Phi}(0) = 1 - \int_0^{a_0} \mu(\tilde{r}) d\tilde{r}, \tag{49}$$

which allows us to obtain

$$\lim_{r\to\infty} m_I(r) = n + \left(\frac{1 - \int_0^{a_0} \mu(\tilde{r}) d\tilde{r} + \xi \sum_{i=1}^{n-1}\left[1 - \int_0^{a_i} \phi(\tilde{r}) d\tilde{r}\right]}{1-\xi}\right). \tag{50}$$

1. Exponential memory with constant background intensity

We now consider, as in Sec. II B 1, the specific case of an exponential memory function along with some constant background intensity [i.e., $\mu(t) = \lambda_0$]. In this case $\phi(t) = \beta e^{-\beta t}$ and therefore $\hat{\phi}(s) = \frac{\beta}{\beta+s}$. Equation (44) in this scenario is given by

$$\hat{\Psi}_i(s) = e^{a_i s}\left[\frac{\beta}{\beta+s} - \beta \int_0^{a_i} e^{-\beta\tilde{r}} e^{-\tilde{r}s} d\tilde{r}\right] \tag{51}$$

$$= \frac{\beta e^{-\beta a_i}}{\beta + s}, \tag{52}$$

and therefore

$$\hat{m}_I(s) = \frac{n}{s} + \frac{1}{s[\beta(1-\xi)+s]}\left[\frac{\lambda_0(\beta+s)}{s} + \xi \sum_{i=1}^{n-1} \beta e^{-\beta a_i}\right]. \tag{53}$$

This allows us to write the following closed-form expression for the expected cascade size at observation time r:

$$m_I(r) = n + \frac{\lambda_0}{1-\xi}\left[r + \frac{\xi}{\beta(1-\xi)}(e^{-\beta(1-\xi)r} - 1)\right] + \sum_{i=1}^{n-1} \frac{\xi e^{-\beta a_i}}{1-\xi}[1 - e^{-\beta(1-\xi)r}]. \tag{54}$$

Note that this equation consists of three components: (i) the number of events we have observed at the start of the prediction window n; (ii) the contribution from the background intensity as we saw in Eq. (24), reflecting the constant background intensity; and (iii) the contribution from the events in the cascade during the observation period. We would like to highlight once more the fact that in the presence of $\lambda_0 \neq 0$, the expected number of events increases linearly with time with a nonzero probability of events occurring at each time step. In contrast, we next consider a decaying background intensity.

2. Exponential background intensity and memory kernel

Suppose that the background intensity is decaying with time such that $\mu(t) = \alpha e^{-\alpha t}$, while the memory kernel is the same as in the previous case, i.e., exponentially distributed with mean time $1/\beta$. Equation (45) is then given by

$$\hat{\Phi}(s) = \frac{\alpha e^{-\alpha(a_0)}}{\alpha + s}, \tag{55}$$

where $a_0 = T$ is the age of the seed comment when the observation period ends. We can obtain the expected number of events at time r in the prediction interval by inverting Eq. (42),

$$m_I(r) = n + \alpha e^{-\alpha a_0}\left\{\frac{1}{\alpha(1-\xi)} + e^{-\alpha r}\left[\frac{\alpha-\beta}{\alpha[(1-\xi)\beta - \alpha]}\right] - e^{-\alpha(1-\xi)r}\left[\frac{\xi}{(1-\xi)(\alpha - [1-\xi]\beta)}\right]\right\} + \sum_{i=1}^{n-1} \frac{\xi\, e^{-\beta a_i}}{1-\xi}[1 - e^{-\beta(1-\xi)r}]. \tag{56}$$

Note that, unlike previous examples such as Eq. (54), this quantity has a finite limit as $r \to \infty$; this is to be expected as there is no longer a constant probability of an event occurring in each time step but rather a decaying probability and as such the likelihood of events occurring after long interevent times is significantly reduced. Thus, we may calculate the large-r expected number of events to be

$$m_I(\infty) = n + \frac{1}{1-\xi}\left[e^{-\alpha a_0} + \sum_{i=1}^{n-1} e^{-\beta a_i}\right]. \tag{57}$$

B. Probability a discussion receives no further comments

In a similar fashion as in Sec. II B 2, we can calculate the probability that the discussion tree receives no further comments in the prediction window by time r, $q_n(r)$, as follows:

$$q_n(r) = \lim_{x\to 0} \frac{I(\{a_0, a_1, \ldots, a_{n-1}\}, r; x)}{x^n}. \tag{58}$$

Since $q_m(r) = 0\,\forall\, m < n$, this results in

$$q_n(r) = \exp\left\{-\left[\int_0^r \mu(r + a_0 - w) + \xi \sum_{i=1}^{n-1} \phi(r + a_i - w) dw\right]\right\}. \tag{59}$$

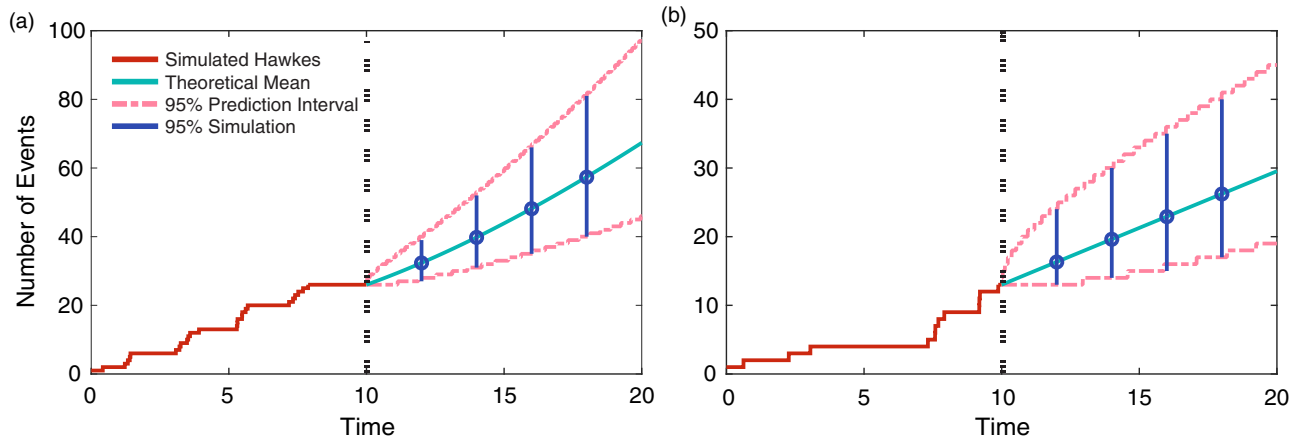

FIG. 6. Numerical simulation and theoretical results for two Hawkes processes where one realization is obtained over the interval [0,10] before proceeding to simulate 10^6 ensembles based upon these events over the interval (10,20]. We show the average number of events along with the 95% percentiles from the distribution of events at a number of time points. In addition, we show the theoretical expected number of events and 95% prediction intervals obtained from numerical inversion of Eq. (38). (a) Case of background intensity $\lambda_0 = 1.5$, fitness $\xi = 0.8$, and exponential memory kernel with mean time 3 ($\beta = 1/3$). (b) Case with shifted power-law memory kernel ($\beta = 1$, $c = 0.01$, $\xi = 0.5$, and $\lambda_0 = 1$).

We can again calculate this quantity exactly in the case of exponential memory with constant background intensity to obtain

$$q_n(r) = \exp\left\{-\left[\lambda_0 r + \xi \sum_{i=1}^{n-1} e^{-\beta a_i}(1 + e^{-\beta r})\right]\right\}. \quad (60)$$

Note that whereas in Eq. (30) the probability of receiving no more comments depended only on exogenous events, in this expression we explicitly account for the endogenous events in the history of the process prior to time T.

IV. COMPARISON OF NUMERICAL SIMULATIONS WITH THEORY

In order to validate these expressions, as well as to test their predictive power, we now proceed to compare them with numerical simulations of Hawkes processes. In each simulation we observe all events that took place before a certain time T, i.e., at $\{\tau_0, \tau_1, \ldots, \tau_{n-1}\}$ with $\tau_{n-1} < T$. Then we predict the evolution of the discussion tree until time Ω, using the PGF describing the distribution of future events from Eq. (38), which we again invert via fast Fourier transforms as in Sec. II C.

Figure 6 shows the case of a Hawkes processes with constant background intensity and an exponential memory kernel [Fig. 6(a)] or a power-law memory kernel [Fig. 6(b)]. In each case, we observe a time series of events until time $T = 10$. Then we make a prediction of the dynamics until time $\Omega = 20$. Specifically, in each panel, we calculate the ensemble average as well as the 95th percentile range, at each time step over all ensembles. This is compared with the theoretical expected number of events as well as the 95% prediction interval with the information available up to time T obtained via numeric inversion of Eq. (38) via the fast Fourier transform method described in Sec. II B 3.

The first thing we observe is that, during the observation window, in the exponential case the frequency of events is much more stable over time in comparison to the bursty nature of the power-law memory kernel, which is characterized by larger periods of inactivity followed by a number of events over a short time interval [65,66]. We also point out the difference between the prediction intervals between the two distributions. Nevertheless, the agreement in both cases between the theoretical predictions and the numerical simulations is exceptionally good.

To determine how the age of comments at time T affects the future popularity of the post, we generate two synthetic time series of events for the observation period, each containing ten events. Then we simulate the rest of the process with those series acting as the history of the content and predict the evolution using our analytical tools with results shown in Fig. 7. In Fig. 7(a) the synthetic series consists of equally spaced interevent times at $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$, while the second sequence shown in Fig. 7(b) is such that there are also ten events but some comments are "younger" at time T, $\{0, 1, 2, 3, 5, 8, 9, 9.5, 9.6, 9.8\}$. Between times T and $\Omega = 20$ we then proceed to simulate the Hawkes process given by Eq. (22) with parameters $\beta = 1/3$, $\lambda_0 = 0.1$, and $\xi = 0.8$. In both cases, we show the theoretical mean given by Eq. (54), the 95% prediction intervals drawn from the distribution at each time step, and also the mean measured directly from the distribution obtained numerically rather than the analytical equation. As we can see, the occurrence of younger posts when making predictions of future popularity results in larger predicted cascade sizes, correctly matched with the results obtained from the simulations both for the expected value and for the upper limits of the prediction intervals.

Also shown in Figs. 7(c) and 7(d) are the corresponding CCDFs, which are the probability of having m or more comments on the discussion tree, at different times. Again, as expected, the larger the age of a given tree, the more likely it is to have received more posts. Also, the process with younger comments by the end of the observation time has higher probability of having more popular discussions due to the time-decaying memory kernel.

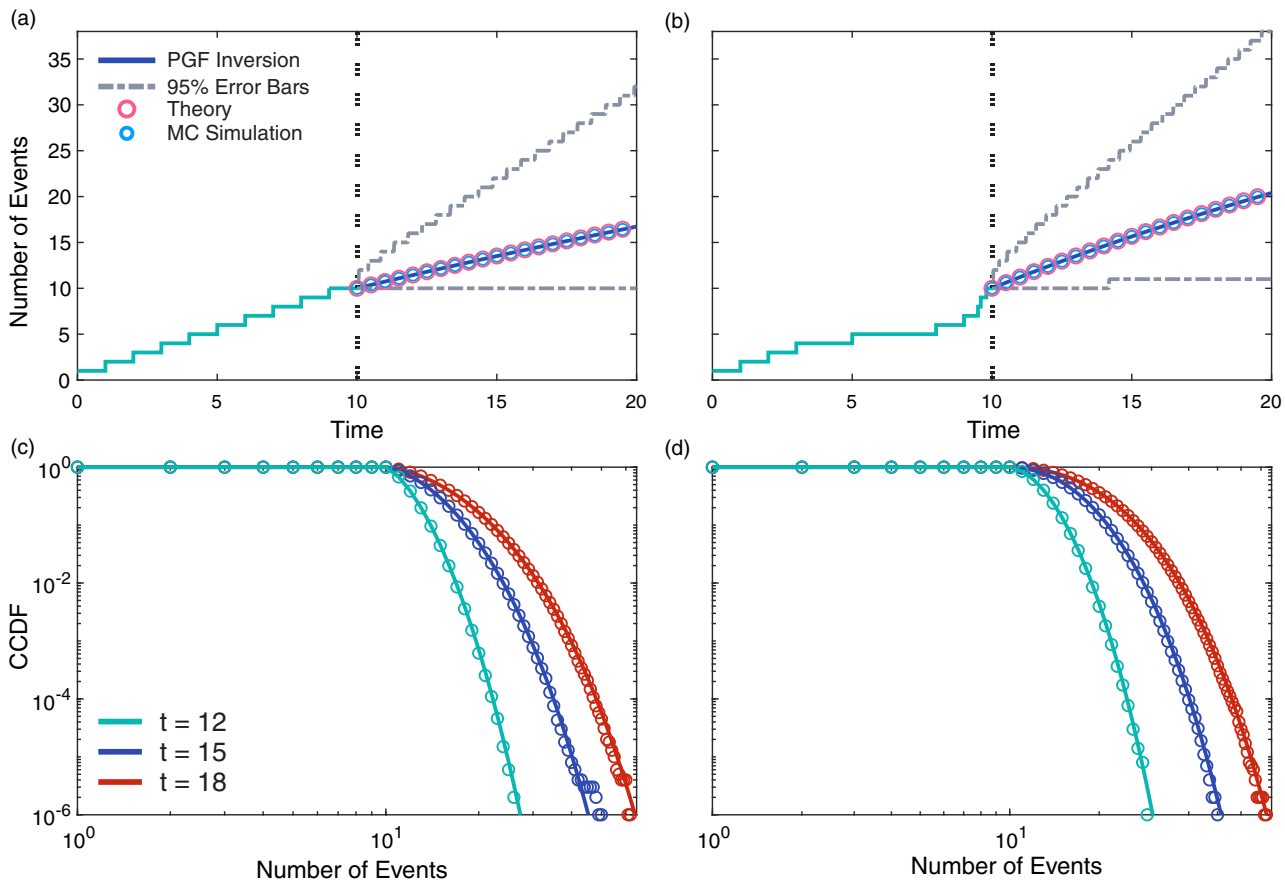

FIG. 7. Comparison of numerical simulations of two half-synthetic Hawkes processes with analytical results. In both cases, the constant background activity is $\lambda_0 = 0.1$ and the excitation function is given by an exponential distribution with mean time 3 ($\beta = 1/3$) and fitness $\xi = 0.8$. Moreover, in both cases, we observe $n = 10$ events before T, although in this case these events are generated synthetically so that (a) $\tau = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ and (b) $\tau = \{0, 1, 2, 3, 5, 8, 9, 9.5, 9.6, 9.8\}$. Then the rest of the process is numerically computed and the prediction of future events is done analytically up to time $\Omega = 20$. The mean number of events is shown in (a) and (b) for the different observation times. In (a) and (b) the circles represent the average over 10^6 Monte Carlo simulations, while the blue line shows the theoretical mean calculated directly from the inversion of the PGF given by Eq. (53). Note that for this particular case of a Hawkes process we may also obtain an analytical expression for this average through Eq. (54). We also show the 95% prediction intervals for the popularities at each time point with the gray lines, again via inversion of Eq. (53). (c) and (d) The corresponding CCDFs for the number of events at a number of different time points are shown, where the lines are the theoretical distributions and the circles are observed CCDFs from the numerical simulations.

Finally, in Fig. 8 we again consider the evenly spaced ten events prior to time $T = 10$ as per Fig. 7(a). In Fig. 8(a) we explore the effect of different fitness parameters ($\xi = 0.2, 0.5, 0.8$) for the discussion's future popularity, while in Fig. 8(b) we show the probability $q_n(r)$, for each of the above fitness values, that the tree does not receive another comment between T and Ω, as given by Eq. (60). The agreement between theory and numerical simulations is again excellent.

V. CONCLUSION

In this paper we have introduced an analytically tractable predictive model built upon a branching process interpretation of the most general form of the Hawkes process, namely, where both the background intensity and the memory kernel are arbitrary time-dependent functions. The focus on the underlying branching process model is important for the generality of the method: While we have considered the dynamics of online opinion boards when describing this model, the bounds on predictability given by the theory are applicable to a wide range of phenomena. Examples of phenomena that can be described by such processes include (but are not limited to) natural catastrophes such as earthquakes [41–43,54,67], the popularity of Youtube videos [60], criminal behavior [68], and conversation frequency [56]. In comparison to related works, a crucial feature of our analysis is that not only does it give the expected number of future events but it also (in a mathematically consistent way) bounds the predictability in the sense of deriving the entire probability distribution of future events. In applications, this allows the direct calculation of not only the expected number of events but also the entire theoretical distribution, without the need for extensive Monte Carlo simulations to determine the time evolution of the process. In turn, this opens the path to the development of, possibly unsupervised, methods that could evaluate and predict expected values (and their confidence intervals)

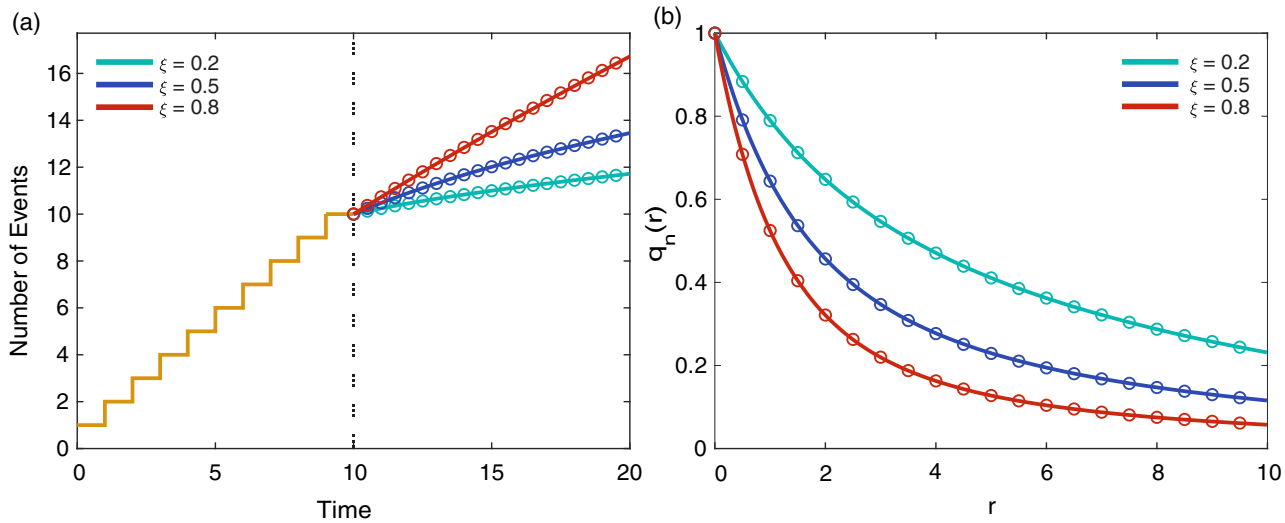

FIG. 8. Simulations equivalent to the one in Fig. 7(a), i.e., with a synthetic evenly spaced series of events before $T = 10$, but with different fitness values for the rest of the evolution up to time Ω. (a) Observed series along with the mean number of events during the prediction window. (b) Probability of zero replies (posts) in a prediction window of length r, $q_n(r)$.

for quantities associated with self-exciting dynamics such as cascade sizes.

The strength of this model's predictions is, like that of all models, limited by the accuracy of the parameters estimated from data to describe the process. Issues regarding the estimation of such parameters to describe empirical data are well known and have been highlighted along with potential remedies in numerous pieces of literature, for example, [22,45,69]. In order to apply the theory introduced in this article to a wide range of applications, we suggest that, while beyond the scope of the present article, future work should focus on improved approaches in estimating parameters, which importantly may in fact be time dependent, as suggested in this article, for both the background intensity and/or the fitness parameter of the memory kernel.

In summary, we believe that the theoretical approach developed in this paper creates an opportunity for more careful prediction of cascading dynamics in online social media. The ability to predict the entire distribution of possible future popularity values means that confidence limits can be placed on predictions, which should allow practitioners to better appreciate the underlying stochasticity of human behavior and to gain some measure of control over its influence on dynamics. Future work will also be devoted to a better characterization of real processes, such as those related to posting on opinion boards, with the aim of learning which kind of background intensity and kernel are best suited to described these processes accurately and how they are correlated with human (bursty) activity.

ACKNOWLEDGMENTS

We thank Kevin Burke, Renaud Lambiotte, and David O'Sullivan for helpful discussions and comments. This work was supported by Science Foundation Ireland Grants No. 16/IA/4470, No. 16/RC/3918, No. 12/RC/2289P2, and No. 18/CRT/6049 (J.D.O'B. and J.P.G.). J.D.O'B. acknowledges support from the Mathematics for Industry Network (Grant No. ECOST/STSM/TD1409/290216/071429). A.A. acknowledges the support of the FPI doctoral fellowship from MINECO (Grant No. FIS2014-55867-P). A.A. and Y.M. also acknowledge partial support from the Government of Aragon, Spain through Grant No. E36-17R (FENOL) and from MINECO and FEDER funds (Grant No. FIS2017-87519-P). A.A. and Y.M. also acknowledges support from Intesa Sanpaolo Innovation Center. We acknowledge the DJEI/DES/SFI/HEA Irish Centre for High-End Computing for the provision of computational facilities and support.

APPENDIX: DERIVATION OF THE BELLMAN-HARRIS INTEGRAL EQUATION

In this Appendix we show how the integral equation of the well-known Bellman-Harris process [32,50,51] may be derived via our differential equation branching process approach. As in the classical discrete-time Galton-Watson model, a single particle is born at time $t = 0$ but rather than living for a single time unit, the particles live until an age a, which is a random variable with distribution $\gamma(a)$. From this definition we may also determine $\lambda(a)$ such that $\lambda(a)\Delta t$ describes the probability that a particle alive at age a dies in the age interval $(a, a + \Delta t)$, or the so-called hazard rate. At the moment of death, the particle produces k offspring or progeny with probability p_k each of which behaves exactly as their parent, i.e., each particle is independent and identically distributed. In the following, we will use the random variable k through its PGF

$$f(x) = \sum_{k=0}^{\infty} p_k x^k. \tag{A1}$$

Due to the previous considerations, these processes are known as age-dependent continuous-time branching processes.

We proceed, as in the main text, to define $G(\tau, a, \Omega; x)$ to be the PGF for the distribution of the number of particles, observed at time Ω, that have lived as a result of a parent of age a producing offspring at time τ. Note that we previously

described this as the size of a tree. However, in the specific case of the Bellman-Harris process, every particle dies at the time of seeding a new tree. Now consider how this PGF may change over infinitesimal time interval $(\tau - \Delta t, \tau)$, where Δt is sufficiently small to ensure only one event at most may occur in this interval. There are two possibilities.

(i) The parent may die with probability $\lambda(a)\Delta t$ producing a random number k children, each of which may start their own subtree at time τ but with age $a = 0$. As we are concerned with the total number of particles to have lived in this interval we must also count the (now deceased) parent particle through an additional power of x in the PGF and as such the total contribution to $G(\tau - \Delta t, a - \Delta t, \Omega; x)$ is $xf[G(\tau, 0, \Omega; x)]$.

(ii) The parent may also live, which occurs with probability $1 - \lambda(a)\Delta t$, and the contribution to $G(\tau - \Delta t, a - \Delta t, \Omega; x)$ in this case is simply $G(\tau, a, \Omega; x)$.

Based on the above, the equation governing the change in G over the time interval $(\tau - \Delta t, \tau)$ is

$$\begin{aligned} &G(\tau - \Delta t, a - \Delta t, \Omega; x) \\ &\quad = \lambda(a)xf[G(\tau, 0, \Omega; x)]\Delta t \\ &\qquad + [1 - \lambda(a)\Delta t]G(\tau, a, \Omega; x), \end{aligned} \tag{A2}$$

which when taking the $\Delta t \to 0$ limit and writing $G(\tau, a, \Omega; x)$ as G becomes

$$\frac{\partial G}{\partial \tau} + \frac{\partial G}{\partial a} = \lambda(a)\{G - xf[G(\tau, 0, \Omega; x)]\}. \tag{A3}$$

Now, as in the main text, we use the method of characteristics with $c = \tau - a$ (being a constant along a characteristic) to express the change in G along a characteristic as

$$\frac{dG}{d\tau} = \lambda(\tau - c)\{G - xf[G(\tau, 0, \Omega; x)]\}, \tag{A4}$$

which, along with the final condition $G(\Omega, a, \Omega; x) = x$, i.e., a tree seeded at observation time has size one, may be solved to give

$$\begin{aligned} G(\tau, a, \Omega; x) &= x \exp\left[-\int_\tau^\Omega \lambda(w - \tau + a)dw\right] \\ &\quad + x\int_\tau^\Omega \exp\left[-\int_\tau^{t'} \lambda(w - \tau + a)dw\right] \\ &\quad \times \lambda(t' - \tau + a)f[G(t', 0, \Omega; x)]dt'. \end{aligned} \tag{A5}$$

Introducing the tree age $t = \Omega - \tau$, i.e., the time between a tree being seeded and observed, and expressing $G(\tau, a, \Omega, x)$ as $G(\Omega - \tau, a; x)$, the above equation becomes

$$\begin{aligned} G(t, a; x) &= x \exp\left[-\int_0^t \lambda(y + a)dy\right] \\ &\quad + x\int_0^t \exp\left[-\int_0^s \lambda(u + a)du\right] \\ &\quad \times \lambda(s + a)f[G(t - s, 0; x)]ds, \end{aligned} \tag{A6}$$

where $y = w - \tau$ and $s = t' - \tau$. The above equation describes the PGF for the size of a tree with tree age t seeded by a parent of age a. Finally, as we are concerned with the total number of particles in a given process which starts through a single parent being born we set $a = 0$ (and drop the $a = 0$ argument from G) to obtain

$$\begin{aligned} G(t; x) &= x \exp\left[-\int_0^t \lambda(y)dy\right] + x\int_0^t \exp\left[-\int_0^s \lambda(u)du\right] \\ &\quad \times \lambda(s)f[G(t - s; x)]ds, \end{aligned} \tag{A7}$$

which we may express as

$$G(t; x) = xS(t) + x\int_0^t \gamma(s)f[G(t - s; x)]ds, \tag{A8}$$

where $\gamma(a)$ describes the probability of living until age a while $S(a) = \int_a^\infty \gamma(s)ds$ gives the probability of a particle having a lifetime of length a or longer, which is the classical Bellman-Harris integral equation.

On a final note, we mention that from Eq. (A8) we may derive the governing equations for two more specific branching processes. First, we obtain the Markovian age-dependent branching process equation through letting $\lambda(a) = \lambda$, i.e., a constant probability of death in each time interval. Second, we can describe the classical Galton-Watson branching process by ensuring each particle has a lifetime of 1, i.e., $\gamma(a) = \delta_{a,1}$, a Dirac delta function with mass at $a = 1$.

[1] P. Lorenz-Spreen, B. Mønsted, P. Hövel, and S. Lehmann, Accelerating dynamics of collective attention, Nat. Commun. **10**, 1759 (2019).

[2] K. E. Matsa and E. Shearer, News use across social media platforms 2018, https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/

[3] H. Kwak, C. Lee, H. Park, and S. Moon, *Proceedings of the 19th International Conference on World Wide Web* (ACM, New York, 2010), pp. 591–600.

[4] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, in *Dynamics On and Of Complex Networks III: Machine Learning and Statistical Physics Approaches*, edited by F. Ghanbarnejad, R. S. Roy, F. Karimi, J.-C. Delvenne, and B. Mitra (Springer Nature, Cham, 2017), pp. 183–204.

[5] P. Aragón, V. Gómez, and A. Kaltenbrunner, *Proceedings of the Eleventh International AAAI Conference on Web and Social Media, Québec, 2017* (AAAI, Palo Alto, 2017).

[6] S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online, Science **359**, 1146 (2018).

[7] S. Goel, A. Anderson, J. Hofman, and D. J. Watts, The structural virality of online diffusion, Manag. Sci. **62**, 180 (2016).

[8] K. Lerman and R. Ghosh, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC, 2010* (AAAI, Menlo Park, 2010).

[9] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, *Proceedings of the 23rd International Conference on World Wide Web* (ACM, New York, 2014), pp. 925–936.

[10] J. M. Miotto and E. G. Altmann, Predictability of extreme events in social media, PLoS ONE **9**, e111506 (2014).

[11] L. Weng, F. Menczer, and Y.-Y. Ahn, Virality prediction and community structure in social networks, Sci. Rep. **3**, 2522 (2013).

[12] R. Bandari, S. Asur, and B. A. Huberman, *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, Dublin, 2012* (AAAI, Palo Alto, 2012).

[13] G. Szabo and B. A. Huberman, Predicting the popularity of online content, Commun. ACM **53**, 80 (2010).

[14] S. Mishra, M.-A. Rizoiu, and L. Xie, *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (ACM, New York, 2016), pp. 1069–1078.

[15] J. P. Gleeson, J. A. Ward, K. P. O'Sullivan, and W. T. Lee, Competition-Induced Criticality in a Model of Meme Popularity, Phys. Rev. Lett. **112**, 048701 (2014).

[16] J. P. Gleeson, K. P. O'Sullivan, R. A. Baños, and Y. Moreno, Effects of Network Structure, Competition and Memory Time on Social Spreading Phenomena, Phys. Rev. X **6**, 021019 (2016).

[17] J. D. O'Brien, I. K. Dassios, and J. P. Gleeson, Spreading of memes on multiplex networks, New J. Phys. **21**, 025001 (2019).

[18] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, Competition among memes in a world with limited attention, Sci. Rep. **2**, 335 (2012).

[19] B. Morrissey, Clients try to manipulate 'unpredictable' viral buzz, Adweek, 19 March 2007.

[20] R. Van der Lans and G. van Bruggen, in *The Connected Customer: The Changing Nature of Consumer and Business Markets*, edited by S. Wuyts, M. G. Dekimpe, E. Gijsbrechts, and R. Pieters (Routledge, New York, 2010), pp. 257–282.

[21] R. Van der Lans, G. Van Bruggen, J. Eliashberg, and B. Wierenga, A viral branching model for predicting the spread of electronic word of mouth, Market. Sci. **29**, 348 (2010).

[22] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, 2nd ed., Probability and Its Applications (Springer, New York, 2003).

[23] K. Lerman, R. Ghosh, and T. Surachawala, Social contagion: An empirical study of information spread on Digg and Twitter follower graphs, arXiv:1202.3162.

[24] P. A. Dow, L. A. Adamic, and A. Friggeri, *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, 2013* (AAAI, Palo Alto, 2013).

[25] A. N. Medvedev, J.-C. Delvenne, and R. Lambiotte, Modelling structure and predicting dynamics of discussion threads in online boards, J. Complex Netw. **7**, 67 (2019).

[26] A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika **58**, 83 (1971).

[27] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2015), pp. 1513–1522.

[28] R. Kobayashi and R. Lambiotte, *Proceedings of the Tenth International AAAI Conference on Web and Social Media, Cologne, 2016* (AAAI, Palo Alto, 2016).

[29] R. Krohn and T. Weninger, Modelling online comment threads from their start, arXiv:1910.08575.

[30] M.-A. Rizoiu, L. Xie, S. Sanner, M. Cebrian, H. Yu, and P. Van Hentenryck, *Proceedings of the 26th International Conference on World Wide Web* (ACM, New York, 2017), pp. 735–744.

[31] J. L. Iribarren and E. Moro, Impact of Human Activity Patterns on the Dynamics of Information Diffusion, Phys. Rev. Lett. **103**, 038702 (2009).

[32] J. L. Iribarren and Esteban Moro, Branching dynamics of viral information spreading, Phys. Rev. E **84**, 046116 (2011).

[33] C. G. Bowsher, Modelling security market events in continuous time: Intensity based, multivariate point process models, J. Econom. **141**, 876 (2007).

[34] E. Bacry, I. Mastromatteo, and J.-F. Muzy, Hawkes processes in finance, Market Microstruct. Liquidity **1**, 1550005 (2015).

[35] E. Bacry and J.-F. Muzy, Hawkes model for price and trades high-frequency dynamics, Quant. Finance **14**, 1147 (2014).

[36] P. Embrechts, T. Liniger, and L. Lin, Multivariate Hawkes processes: An application to financial data, J. Appl. Probab. **48**, 367 (2011).

[37] M. Rambaldi, P. Pennesi, and F. Lillo, Modeling foreign exchange market activity around macroeconomic news: Hawkes-process approach, Phys. Rev. E **91**, 012819 (2015).

[38] Y. Aït-Sahalia, J. Cacho-Diaz, and R. J. A. Laeven, Modeling financial contagion using mutually exciting jump processes, J. Financ. Econ. **117**, 585 (2015).

[39] V. Pernice, B. Staude, S. Cardanobile, and S. Rotter, How structure determines correlations in neuronal networks, PLoS Comput. Biol. **7**, e1002059 (2011).

[40] F. Gerhard, M. Deger, and W. Truccolo, On the stability and dynamics of stochastic spiking neuron models: Nonlinear Hawkes process and point process GLMs, PLoS Comput. Biol. **13**, e1005390 (2017).

[41] Y. Ogata, Statistical models for earthquake occurrences and residual analysis for point processes, J. Am. Stat. Assoc. **83**, 9 (1988).

[42] Y. Ogata, Space-time point-process models for earthquake occurrences, Ann. Inst. Stat. Math. **50**, 379 (1998).

[43] A. Saichev and D. Sornette, Universal Distribution of Interearthquake Times Explained, Phys. Rev. Lett. **97**, 078501 (2006).

[44] R. Shcherbakov, J. Zhuang, G. Zöller, and Y. Ogata, Forecasting the magnitude of the largest expected earthquake, Nat. Commun. **10**, 1 (2019).

[45] M.-A. Rizoiu, Y. Lee, S. Mishra, and L. Xie, *Frontiers of Multimedia Research* (ACM, New York, 2017), pp. 191–218.

[46] A. Saichev and D. Sornette, Fertility heterogeneity as a mechanism for power law distributions of recurrence times, Phys. Rev. E **87**, 022815 (2013).

[47] T. Hogg and K. Lerman, Social dynamics of Digg, EPJ Data Sci. **1**, 5 (2012).

[48] F. Wu and B. A. Huberman, Novelty and collective attention, Proc. Natl. Acad. Sci. USA **104**, 17599 (2007).

[49] H. S. Wilf, *Generatingfunctionology* (Elsevier, New York, 2013).

[50] K. B. Athreya and P. E. Ney, *Branching Processes* (Springer Science + Business Media, New York, 2013).

[51] T. E. Harris, *The Theory of Branching Processes* (Courier, Chelmsford, 2002).

[52] R. Haberman, *Elementary Applied Partial Differential Equations* (Prentice Hall, Englewood Cliffs, 1983), Vol. 987.

[53] A. G. Hawkes and D. Oakes, A cluster process representation of a self-exciting process, J. Appl. Probab. **11**, 493 (1974).

[54] K. Ogata, *Modern Control Engineering*, 4th ed. (Prentice-Hall, Upper Saddle River, 2002).

[55] E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi, Self-exciting point process models of civilian deaths in Iraq, Security J. **25**, 244 (2012).

[56] N. Masuda, T. Takaguchi, N. Sato, and K. Yano, in Temporal Networks, edited by P. Holme and J. Saramäki (Springer Nature, Cham, 2013), pp. 245–264.

[57] T. Aoki, T. Takaguchi, R. Kobayashi, and R. Lambiotte, Input-output relationship in social communications characterized by spike train analysis, Phys. Rev. E **94**, 042313 (2016).

[58] D. Oakes, The Markovian self-exciting process, J. Appl. Probab. **12**, 69 (1975).

[59] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, Endogenous Versus Exogenous Shocks in Complex Networks: An Empirical Test Using Book Sale Rankings, Phys. Rev. Lett. **93**, 228701 (2004).

[60] R. Crane and D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proc. Natl. Acad. Sci. USA **105**, 15649 (2008).

[61] H.-H. Jo, J. I. Perotti, K. Kaski, and J. Kertész, Correlated bursts and the role of memory range, Phys. Rev. E **92**, 022814 (2015).

[62] J. K. Cavers, On the fast Fourier transform inversion of probability generating functions, IMA J. Appl. Math. **22**, 275 (1978).

[63] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, Queueing Syst. **10**, 5 (1992).

[64] P. G. Fennell, S. Melnik, and J. P. Gleeson, Limitations of discrete-time approaches to continuous-time contagion dynamics, Phys. Rev. E **94**, 052125 (2016).

[65] A.-L. Barabasi, The origin of bursts and heavy tails in human dynamics, Nature (London) **435**, 207 (2005).

[66] M. Karsai, H.-H. Jo, and K. Kaski, *Bursty Human Dynamics* (Springer, Berlin, 2018).

[67] A. Helmstetter and D. Sornette, Subcritical and supercritical regimes in epidemic models of earthquake aftershocks, J. Geophys. Res. **107**, 10 (2002).

[68] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, Self-exciting point process modeling of crime, J. Am. Stat. Assoc. **106**, 100 (2011).

[69] J. G. Rasmussen, Bayesian inference for Hawkes processes, Methodol. Comput. Appl. Probab. **15**, 623 (2013).