Hawkes process as a model of social interactions: a view on video dynamics

Hawkes process as a model of social interactions: a view on video dynamics

Lawrence Mitchell[1,2] and Michael E Cates[1]

[1] SUPA, School of Physics and Astronomy, University of Edinburgh, JCMB Kings Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK
[2] EPCC, University of Edinburgh, JCMB Kings Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK

E-mail: lawrence.mitchell@ed.ac.uk

Abstract
We study by computer simulation the 'Hawkes process' that was proposed in a recent paper by Crane and Sornette (2008 *Proc. Natl Acad. Sci. USA* **105** 15649) as a plausible model for the dynamics of YouTube video viewing numbers. We test the claims made there that robust identification is possible for classes of dynamic response following activity bursts. Our simulated time series for the Hawkes process indeed fall into the different categories predicted by Crane and Sornette. However, the Hawkes process gives a much narrower spread of decay exponents than the YouTube data, suggesting limits to the universality of the Hawkes-based analysis.

PACS numbers: 89.75.Fb, 89.20.Hh, 05.40.−a

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Recently [1], Crane and Sornette analysed the viewing of YouTube videos as an example of a nonlinear social system. They identified peaks in the time series of viewing figures for around half a million videos and studied the subsequent decay of the peak to a background viewing level. A self-excited Poisson process, or Hawkes process [2], was proposed as a model of the video-watching dynamics, and a plausible link made to the social interactions that create strong correlations between the viewing actions of different people. Individual viewing is not random but influenced by various channels of communication about what to watch next.

The Hawkes process has a Poisson distributed number of views, with an instantaneous rate given by

$$\lambda(t) = \eta(t) + \sum_{\{t_i < t\}} \mu_i \phi(t - t_i). \tag{1}$$

Here η is a noisy source term (allowing viewing to occur even for a completely dormant video, for instance) and the summation describes how past viewing events at times $\{t_i\}$ influence the current event rate. The coefficient μ_i is the number of potential viewers influenced directly by person i who viewed a video at time t_i; the function ϕ describes the waiting time distribution for those influenced, between trigger and response. (Put differently, this is the distribution of waiting times between finding out about a particular video and actually viewing the video.) If $\phi(t)$ is a power-law memory function as used in this work, the resulting process is also known as an epidemic type aftershock sequence (ETAS) model [3], and can be used to model earthquake aftershocks.

On the basis of previous work [4], Crane and Sornette chose a long-memory waiting time distribution

$$\phi(t) \sim t^{-(1+\theta)}, \qquad \theta \in (0, 1). \tag{2}$$

For fixed θ (we address variability in θ later on), the behaviour of a time series generated by such a Hawkes process then depends on the distribution of μ. There are four separate dynamical classes, two if a viewing shock happens from an external stimulus, two from internal dynamics [5]. In this paper we address only the dynamics of externally shocked time series, but for completeness all four dynamic classes will be outlined below.

In each dynamic class, there is a different prediction for the power-law decay of the activity level following an initial shock. The power laws involved are quite distinct for each class, and predicted by [5] to depend solely on θ, whereas the statistics of μ control solely which class one is in. Therefore, if equation (2) were really to hold (with a unique θ) one might naively expect the distribution of power-law exponents observed in the data to collapse onto a set of discrete delta-functions, one for each dynamic class. On reflection, however, this cannot be correct since an individual activity burst represents a sequence of discrete events which (unless the total number of these is enormously large) is unlikely to be fully self-averaging for the purposes of fitting a power law to the long-time decay. In practice for the YouTube data [1] the distribution of fitted exponents is very broad with, at best, bump-like features at the expected discrete exponent values. Crane and Sornette get around this by using a quite separate method (detailed below) to classify the dynamic class of each burst, and then showing that the subdistributions for each class are unimodal with modal values close to the expected one for that class. The overlapping exponent distributions that necessitate this procedure do however call into question the announced robustness [1] of the dynamic classes inferred from the Hawkes model.

In the present work we perform simulations that shed light on how much of this exponent variability can be expected to arise from the Hawkes process itself. Any variability beyond this level in the YouTube data is evidence that equation (2) does not really describe the social dynamics of YouTube. Of course, nobody would expect these dynamics to be captured *exactly* by the Hawkes process; however, behind the concept of robust dynamic classes in [1] lies a broader notion of universality. For instance in equilibrium critical phenomena, a very simple model (the Ising model) captures to arbitrary accuracy the universal features of a wide class of phase transitions involving order parameters of the same symmetry. In the wider context of nonequilibrium criticality, the universal status of simple models is less well established, and deserves detailed attention. Our simulation results suggest that this universality may

be somewhat limited, at least if one is interested in the *distribution* of fitted exponents for individual activity bursts within each dynamic class.

In what follows we first classify all four dynamic regimes before presenting the analysis of our results.

2. External shock

In this regime, the viewing rate is first dominated by the η term in equation (2). At some time, t_0, the video gains widespread public attention. (It might be featured in a national newspaper, or on some high-traffic website; or it may relate to a famous person whose death is suddenly announced.) This produces a spike in the viewing figures which then decay away. The form of the decay depends on the distribution of μ.

- If $\langle \mu \rangle = 1$, a cascade of viewing events occurs and the time series decays from the shock like $\sim t^{-(1-\theta)}$. This is termed a critical decay.
- If $\langle \mu \rangle < 1$, only first-generation viewing events are important (i.e. those stimulated by the external source) and the time series decays like $\sim t^{-(1+\theta)}$. This is a subcritical decay.

3. Internal shock

A particular series of viewing events can lead to an internally created maximum in the time series (above that expected for a Poisson noise process). This internal shock has a different decay exponent again from the externally induced peaks. The two internal dynamic classes are as follows.

- A simple noise process if $\langle \mu \rangle < 1$; no coherent peak arises.
- A peak grows and decays like $\sim t^{-(1-2\theta)}$; this occurs if $\langle \mu \rangle = 1$.

4. Classification and exponent values

If this model is correct for the dynamics of video views, it should be possible to identify the different dynamic classes by finding peaks in the viewing time series and then fitting a power law to the subsequent decay. These power laws should form a distribution which arises as the merger of the various classes; if the individual activity bursts can separately be classified, the subdistribution for each class can be extracted. Crane and Sornette perform such an analysis and by fitting to the modal exponent for each class infer a value for the exponent θ in equation (2) of $\theta \approx 0.4$. We therefore create artificial time series with $\langle \theta \rangle = 0.4$ for best comparison with their data. With this choice, we expect to extract decay exponents (recalling that we only study the externally shocked case) of

- $\beta_{sc} = 1 + \theta = 1.4$ and
- $\beta_c = 1 - \theta = 0.6$.

As mentioned previously, the model might lead one to expect δ-function peaks in a PDF of decay exponents, corresponding to the various dynamic classes. The data presented in [1] show weak peaks at these values, but with a significant spread. In particular, some of the extracted exponents would imply values of θ that lie outside the range $0 < \theta < 1$ required by the model itself. With the help of our simulation data, we can look at whether the spread arises through miscategorization; a poor fitting method; fluctuations in the fitted exponents due to noise inherent in the Hawkes process itself or failure of the Hawkes model to accurately describe the YouTube data.

5. Generating the synthetic data

We carry out a discrete-time simulation of the Poisson/Hawkes process, restricting attention to activity bursts initiated by external shocks. (We generally take each initial shock to comprise $N_0 = 5000$ views.) We choose to generate a random number of views from a Poisson distribution with given mean at each time step and update the rate accordingly afterwards. Effectively, we treat the continuously varying $\lambda(t)$ as a constant, generate a given number of events and then modify λ according to equation (1). We must also choose what values θ and μ can take. Additionally, we need a normalization for the distribution of waiting times, ϕ. Following [1] we take this distribution normalized to unity (so that all those 'influenced' to watch a video by a particular viewing event do watch it eventually). Remembering that the waiting time will be an integer (due to our simulation strategy), we have

$$\phi(t) = \frac{1}{t^{1+\theta}\zeta(1+\theta)} \tag{3}$$

with ζ the Riemann ζ-function. This ensures

$$\sum_{t=1}^{\infty} \phi(t) = 1. \tag{4}$$

Our algorithm for generating the synthetic data is therefore as follows.

 (i) Shock the system by creating $N(0) = N_0$ initial viewing events.
 (ii) For each viewing event, generate the number of subsequent viewers μ_i by sampling from $P(\mu)$. At time t we therefore seed $n = \sum_{i=1}^{N(t)} \mu_i$ future viewing events. Each of these n future viewers has their own decay constant θ_i drawn from $P(\theta)$.
 (iii) Generate a Poisson event rate $\lambda(t)$ by summing over the past history according to equation (1).
 (iv) Use this rate to generate the number of viewing events $N(t)$ between time t and $t + \delta t$.
 (v) Increment t by δt and repeat steps (ii)–(v) until the maximum specified time has been reached.

The model analysis in [1] assumes that all θ values are equal. It is, however, not clear that all interactions would involve exactly the same response kernel. If those influenced have a distribution of waiting habits, this can safely be averaged unless there is a correlation with the person exerting the influence (so that θ varies between the events i in equation (2)). To allow for the latter possibility we carry out simulations not only with a single $\theta = 0.4$ but with a random distribution of θ_i to see if this modifies the results. For the latter we choose θ from a truncated Gaussian with mean 0.4 and standard deviation 0.2 (restricting the support to $\theta \in (0, 1)$).

Finally we need to choose the statistics of μ_i. We will see that the particular choice of distribution does not make an appreciable difference to the results for the externally shocked system (although as detailed above, the value of $\langle \mu \rangle$ is important). Here we present results where μ is drawn from appropriately weighted δ-function distributions as well as Poisson distributions.

6. Fitting the data

We estimate decay exponents from the artificially produced time series both via the method described in [1] and using a maximum likelihood estimator. The least squares estimator used in [1] can give incorrect parameters [6] since some of the assumptions behind it are violated

for power-law decays. However, in our study we find little difference between the maximum likelihood decay exponents and the least squares exponents, which is evidence that errors in the exponent estimation method used in [1] are not the main cause of the large spread of observed exponents.

6.1. Maximum likelihood estimator

Each post-shock time series decay has two free parameters once the peak has been identified: the decay exponent β, and the time at which the peak has decayed to the background noise level $t_{\max}$. To construct the maximum likelihood estimator (MLE) for β, we assume the data to be independent identically distributed random variables drawn from a discrete power-law distribution, $P(t, \beta)$. That is, with a peak occurring at $t = 0$, we expect, for $t = 1, \ldots, t_{\max}$

$$P(t, \beta) = \frac{t^{-\beta}}{H_{t_{\max}, \beta}} \tag{5}$$

where

$$H_{t_{\max}, \beta} = \sum_{k=1}^{t_{\max}} k^{-\beta} \tag{6}$$

is a generalized harmonic number.

For every possible $t_{\max}$ we find the best fit value of β for this distribution using a maximum likelihood estimator. For a time series $A(t)$, our data set consists of $A(t_i)$ observations at each time t_i. Each of these t_is has an individual likelihood given by $P(t_i, \beta)$. We assume each observation of t_i is independent and so the likelihood of the data set factorizes into the product of the individual likelihoods:

$$\mathcal{L}(\beta) = \prod_{\{t_i \leqslant t_{\max}\}} P(t_i, \beta). \tag{7}$$

To find the best fit value of β for a given data set, we maximize the likelihood with respect to β. (In fact, since the likelihood is such a small value, we instead maximize the logarithm of the likelihood, but this gives the same result.) To find the best value of $t_{\max}$ we follow the method of [6] and choose $t_{\max}$ which minimizes the Kolmogorov–Smirnov distance statistic. That is, for each value of $t_{\max}$ we find the best fit decay exponent and calculate

$$D = \max_x |E(x) - C(x)| \tag{8}$$

where $E(x)$ is the empirical cumulative distribution function, and $C(x)$ is the best fit hypothesis cumulative distribution function. We then pick $t_{\max}$ as that value which minimizes D. (Note that this fitting method, while finding the best fit, tells us nothing about the quality of that fit.)

6.2. Least squares estimator

We also calculate the decay exponents for the same time series using the method described by Crane and Sornette [1]. This uses a least squares regression on the data set to find the decay exponent from the peak over a fixed window size. For each fit, they look at the distribution of the relative residuals, i.e. the difference between the model and the empirical data, divided by the expected value at that point. If the relative residuals are not distributed normally, the fit for that window size is rejected. The best fit to the data is chosen to be the largest window size with normally distributed relative residuals. Following [1] we reject the fit if the hypothesis of a normal distribution is violated at the 1% level using a χ^2 test.

6.3. Fitting to an ensemble average

The individual time series that we generate are subject to a reasonable amount of noise giving a spread of best fit decay exponents. Given that we control the time and size of the initial shock, we can easily obtain better statistics for the different parameter choices by considering ensemble averaged time series. This allows us to observe the behaviour of the decay for a longer time and get a better fit for the decay exponents.

The fitting method in this case is as before; we obtain the best fit β value by maximizing the likelihood of the data. The decay now occurs over the whole tail of the time series and so we do not need to find t_{max}; we can set it manually as equal to the largest time in our data set. We fit both from the peak of the decay and the 'tail'. To determine where the tail of the data starts, we follow the same procedure as detailed above for finding t_{max}, only this time we apply it to find t_{min}. That is, for each t_{min} value, we calculate D (equation (8)) of the best fit and subsequently choose as our lower cut-off that t_{min} which minimizes D. We obtain errors on our estimates of β by noting that our MLE is asymptotically optimal [7]; for N observations, the variance in the estimated value is therefore given by the inverse of the observed Fisher information [8]

$$\mathrm{Var}(\beta) = \frac{1}{N\,\mathcal{J}(\beta)} \tag{9}$$

with

$$\mathcal{J}(\beta) = -\frac{1}{N}\frac{\partial^2 \log \mathcal{L}(\beta)}{\partial\beta^2} \tag{10}$$

which can easily be obtained numerically. The MLE is asymptotically Gaussian with mean β and variance given by equation (9) and so confidence intervals are just the standard Gaussian ones.

7. Results

To begin, we look at the behaviour of the ensembled-averaged time series. As expected, we see a clear distinction between the subcritical decay (where $\langle\mu\rangle < 1$) and the critical decay (with $\langle\mu\rangle = 1$). The best fit decay exponents are also those expected (figure 1).

The difference between critical and subcritical decays remains when we draw θ from a Gaussian distribution (figure 2). We do, however, note a significant difference from the single-valued θ case: the numerical values of the decay exponents no longer agree well with the predicted values.

Note how the subcritical decay appears to exhibit a crossover between a short-time 'critical' decay exponent and long-time subcritical decay. Increasing $\langle\mu\rangle$ towards the critical value of unity moves the crossover to later and later times. Interestingly, when θ is drawn from a Gaussian distribution, both the critical and subcritical decays exhibit some sort of crossover behaviour, not seen in the critical decay for single-valued θ. This crossover can contribute to the spread of exponents in the subcritical case, since the fitting mechanism may pick up the early time decay. The crossover observed is discussed in detail in the context of the ETAS model in [3, 9].

We now look at the distribution of decay exponents of individual time series obtained from both the MLE and least squares estimator. Our results for the ensembled-averaged time series indicate that we will likely not see the asymptotic exponent in the subcritical case if we fit the entire post-shock time series ($t_{min} = 0$), as the decay exponent will be skewed by the early time 'critical' decay. We therefore show results with $t_{min} = 0, 10, 20$ and 30; these
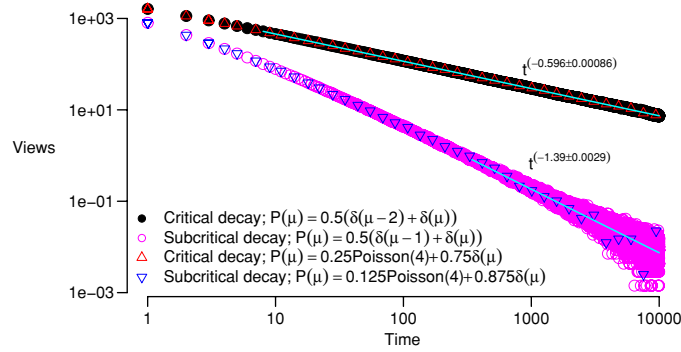
Figure 1. Ensemble average decay exponents with $P(\theta) = \delta(\theta - 0.4)$ and $P(\mu)$ as indicated. Each data set is the average of 700 realizations with an initial shock of 5000 views. Lines show the best fit decay exponent in the tail of the decay ($t \geqslant 8$ for critical and $t \geqslant 344$ for subcritical decay) obtained from the MLE, $\pm$ figures are 95% confidence intervals whose calculation is detailed in the text. The decay exponents are in good agreement with the theoretical values of 0.6 and 1.4.
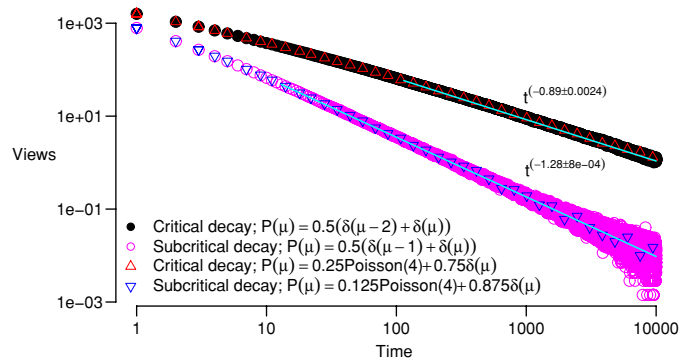

Figure 2. Ensemble average decay exponents with $P(\theta) = N(0.4, 0.2)$ and $P(\mu)$ as indicated. Each data set is the average of 700 realizations with an initial shock of 5000 views. Lines show the best fit decay exponent in the tail of the decay ($t \geqslant 114$ for critical and $t \geqslant 14$ for subcritical decay) obtained from the MLE, $\pm$ figures are 95% confidence intervals. The decay exponents are no longer in good agreement with the theoretical values of 0.6 and 1.4.

latter fits will give us an indication of what the tail exponent looks like. The results for a single value of θ are shown in figure 3, those with θ from a Gaussian distribution are shown in figure 4.

As expected, fitting the entire post-peak time series underestimates the subcritical decay exponent. Both fitting methods pick up the early time decay, which is slower; once the early time peak is ignored, the decay exponents are more similar to the tail seen in the ensemble-averaged case. We note that the fits for $t_{min} \gg 1$ do have quite a large spread of exponents. This is due to poor statistics in the tail of the decay: the fluctuations are large enough that we occasionally pick up a highly anomalous decay exponent. This form of statistical noise appears to be intrinsic to the Hawkes process once the data are filtered by t_{min}. Improved fits, and presumably also narrower distributions of the fitted exponents, would arise if we used much larger initial shocks ($N_0 \gg 5000$).
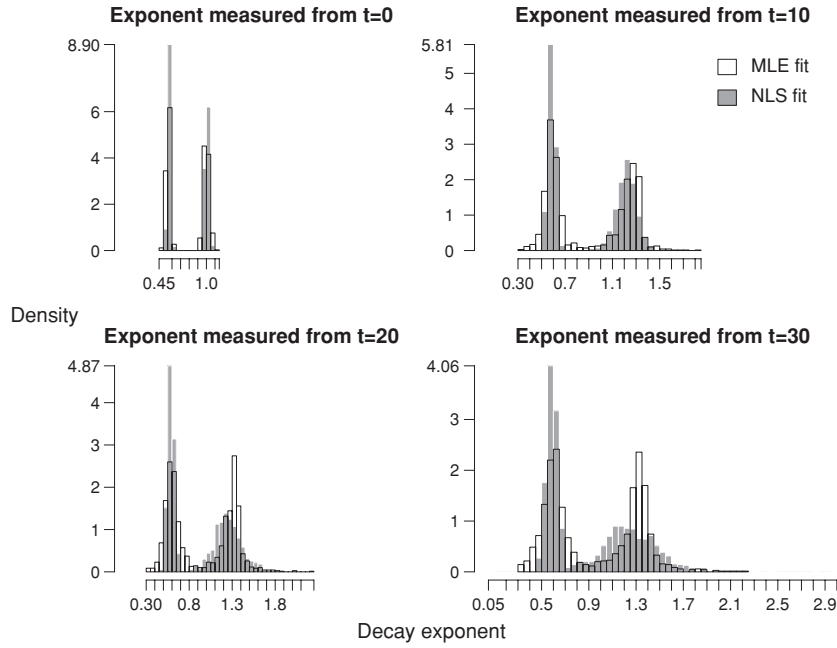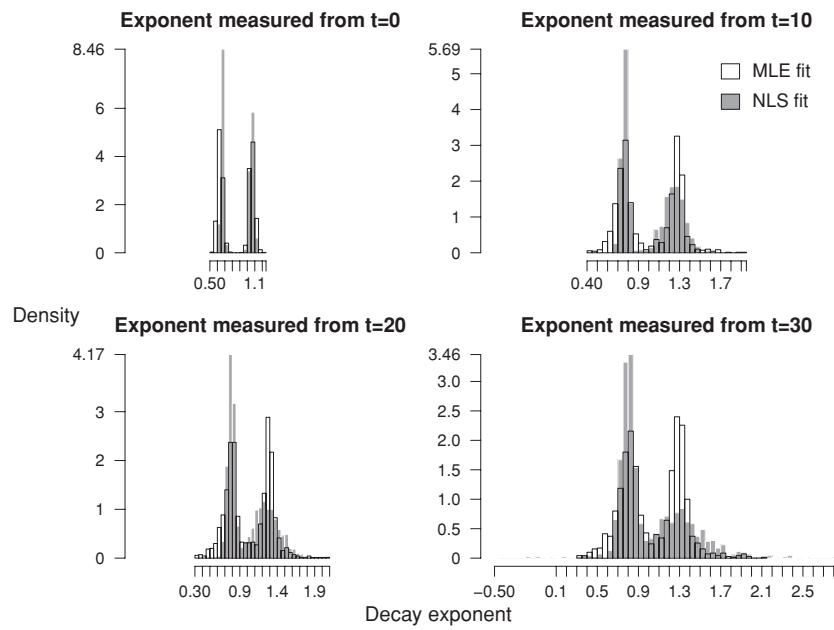
Figure 3. Histograms of extracted decay exponents for critical and subcritical time series and $P(\theta) = \delta(\theta - 0.4)$, initial shock is 5000, 700 realizations. The grey histograms show exponents obtained from nonlinear least squares fitting, white histograms are obtained from MLE fits. Note how there are two distinct peaks in the distribution, corresponding to critical and subcritical decays. The subcritical peak moves from $\beta \approx 1$ to $\beta \approx 1.4$ when we avoid picking up the early time critical decay described in the text.

Our choice of N_0 is however consistent with the statements in [1] of mean total views in the tens of thousands (with at least 20% of these viewed on the peak day) for the shocked case. It is intended to give a realistic estimate of the intrinsic noise in the Hawkes process, to see if this can account for the large exponent spread actually found in [1]. Comparison of their figure 4 with our figures 3 and 4 shows such an explanation to be implausible: the exponent spread in the YouTube data is much too large, particularly for the subcritical case. We have also performed simulations with $N_0 = 500$ and $N_0 = 50\,000$, i.e. one order of magnitude in either direction from the results reported here. In studying the ensemble average of 700 such time series from these simulations, we find that we cannot reject, at the 95% significance level, the hypothesis that the data are the same as those we have reported for $N_0 = 5000$. In other words, the size of the initial shock does not affect the statistics of the resulting time series. For the small initial shocks ($N_0 = 500$), the spread of individually fitted exponents is indeed larger than those we show here with $N_0 = 5000$ and vice versa for the larger shocks ($N_0 = 50\,000$). This is simply due to the fitting method being more (less) affected by statistical fluctuations. The peaks of the exponent distributions do not, however, change appreciably.

7.1. Classifying time series

Crane and Sornette do not have *a priori* knowledge of which dynamic class each time series belongs to. Because the exponents do not fall into clear classes, they use a classification method based on the fraction of the total views that arise on the day of maximal viewing, termed the 'peak fraction' (F). (This fraction is of course a measure of the steepness of the

Figure 4. Histograms of extracted decay exponents for critical and subcritical timeseries and $P(\theta) = N(0.4, 0.2)$, initial shock is 5000, 700 realizations. The grey histograms show exponents obtained from nonlinear least squares fitting, white histograms are obtained from MLE fits. Note how there are two distinct peaks in the distribution, corresponding to critical and subcritical decays. The subcritical peak moves from $\beta \approx 1$ to $\beta \approx 1.3$ when we avoid picking up the early time critical decay described in the text.
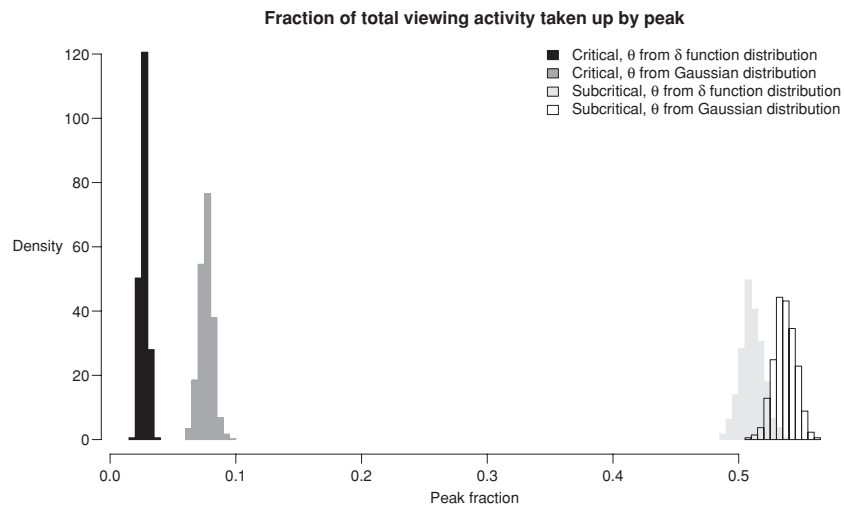
Figure 5. Distribution of peak fractions for critically and subcritically decaying time series. Parameter values as indicated in legend. Choosing a cutoff value of $F = 0.2$ for the peak fraction would result in no misclassifications.

subsequent decay, hence of β.) In our simulations, since we know $\langle \mu \rangle$, and hence which class any given time series is actually in, we can look at the peak fraction and see if this method

9

allows for any misclassification. The classification according to F in [1] is to consider $F \geqslant 0.8$ as an exogenous subcritical decay, $0.2 < F < 0.8$ as exogenous critical decay and $F \leqslant 0.2$ as endogenous critical decay. There are some further comments that the classification between the two exogenous cases is not significantly altered when varying the boundary between $F = 0.7$ and $F \to 1$. We have not calibrated our simulations to any of Crane and Sornette's data, and hence do not know how long the time increment in our updates is relative to their data. The boundaries we choose for classification will therefore not have the same numerical values; this will not, however, invalidate our study of the classification method. We find that our simulated time series show two well-defined peaks in the distribution of peak fractions. Choosing a cut-off of $F < 0.2$ to define exogenous critical and $F > 0.2$ to define exogenous subcritcal decay (recall we do not treat the endogenous case) results in no misclassification.

Figure 5 shows a histogram of peak fractions of simulations with $P(\theta) = \delta(\theta - 0.4)$ and with a Gaussian distribution. In both cases, there is an obvious divide between subcritically decaying time series and critical decays. For a suitably placed boundary between high and low peak fractions ($F = 0.2$), this method correctly classifies every single time series we have studied.

8. Conclusions

The observed behaviour of the Hawkes process subject to external shock is, for the case of a single-valued θ distribution, exactly as predicted in [1, 5]. When θ is drawn from a broad distribution, the numerical values of the decay exponent are modified, but the overall picture of critical and subcritical decays remains. Our results show a significant spread of fitted decay exponents, though much less than that seen in the YouTube data reported in [1]. We can, however, shed some light on this. We have good control of all the time series we fit to; in particular, we ensure that they are all subject to the same size of fluctuation (by always studying the same size of shock). Crane and Sornette do not have this luxury. Our fitting to the tails of individual time series shows that the exponent can vary widely if the statistics are poor (in some instances the best fit exponent is very different from that of the underlying distribution of which a given time series represents a single sample). It seems likely that some of the breadth in the range of exponents seen in [1] is caused by considering many time series with poor statistics in the tail. By only considering time series with particularly large peaks (relative to the background viewing rate), a set of decay exponents with lower variance might be obtained. Of course, this would have a cost in terms of the overall statistics of the sample.

In addition, our study shows that the peak fraction classification method is a good one and we suggest that carrying out this classification and then fitting to the ensemble average of suitably normalized time series may give the best estimate of the decay exponents. We have also shown that the least squares fitting method gives results that are not very different from the maximum likelihood approach favoured here.

Our results demonstrate a way to test if θ is really a unique global constant (equivalently, drawn from a δ-function distribution). The ensemble-averaged time series in this case are measurably different from those where θ is broadly distributed with the same mean. Particularly, we observe a crossover effect in the critical decay for a broad θ distribution that is not present if θ is constant. If the time series can be correctly classified using the peak-fraction method, an ensemble average of (suitably normalized) critical time series might be diagnostic of whether θ is effectively constant or not.

Finally, we reiterate that although our analysis of synthetic Hawkes process data results in a spread of fitted exponents within each dynamic class, this intrinsic noise does not fully account for the much wider distributions seen in the YouTube data of Crane and Sornette

[1]. This suggests limits to the universality concept which presumably underlies attempts to classify activity bursts in social systems into 'robust dynamic classes' [1]. While the Hawkes process is clearly useful in analysing real-world data from complex social systems, some fairly basic observables, such as the variance of the exponent distribution for individual activity bursts, are seemingly not captured by it. These aspects are thus either nonuniversal or lie in the universality class of a more complex model than Hawkes.

Acknowledgments

References

[1] Crane R and Sornette D 2008 Robust dynamic classes revealed by measuring the response function of a social system *PNAS* **105** 15649–53
[2] Hawkes A G 1971 Spectra of some self-exciting and mutually exciting point processes *Biometrika* **58** 83
[3] Helmstetter A and Sornette D 2002 Subcritical and supercritical regimes in epidemic models of earthquake aftershocks *J. Geophys. Res.* **107** 2237
[4] Johansen A and Sornette D 2000 Download relaxation dynamics on the WWW following newspaper publication of URL *Physica* A **276** 338–45
[5] Sornette D and Helmstetter A 2003 Endogenous versus exogenous shocks in systems with memory *Physica* A **318** 577–91
[6] Clauset Aaron, Shalizi Cosma Rohilla and Newman M E J 2007 Power-law distributions in empirical data arXiv:0706.1062 [physics.data-an]
[7] Cox D R and Hinkley D V 1974 *Theoretical Statistics* (London: Chapman and Hall)
[8] Tamhane A C and Dunlop D D 2000 *Statistics and Data Analysis* (New Jersey: Prentice-Hall)
[9] Saichev A and Sornetter D 2005 Vere-Jones' self-similar branching model *Phys. Rev.* E **72** 056122