



Nonparametric self-exciting models for computer network traffic

Matthew Price-Williams¹ · Nicholas A. Heard^{1,2}

Received: 29 June 2018 / Accepted: 17 April 2019
© The Author(s) 2019

Abstract

Connectivity patterns between nodes in a computer network can be interpreted and modelled as point processes where events in a process indicate connections being established for data to be sent along that edge. A model of normal connectivity behaviour can be constructed for each edge in a network by identifying key network user features such as seasonality or self-exciting behaviour, since events typically arise in bursts at particular times of day which may be peculiar to that edge. When monitoring a computer network in real time, unusual patterns of activity against the model of normality could indicate the presence of a malicious actor. A flexible, novel, nonparametric model for the excitation function of a Wold process is proposed for modelling the conditional intensities of network edges. This approach is shown to outperform standard seasonality and self-excitation models in predicting network connections, achieving well-calibrated predictions for event data collected from the computer networks of both Imperial College and Los Alamos National Laboratory.

Keywords Computer network · Wold process · Hawkes process · Changepoint estimation

1 Introduction

Statistical anomaly detection tools (Lazarevic et al. 2003; Neil et al. 2013; Heard and Rubin-Delanchy 2016) have an important role to play in the next generation of cyber-security defences, complementing more traditional *signature-based* techniques which rely on packet inspection to match known indicators of malicious content. In contrast, anomaly detection techniques harvest data on *normal* behaviour in a computer network, and monitor for any significant deviations in observed traffic from probability models built on the historic data.

Good anomaly detection power requires good underlying models, and the focus of this article is on building accurate models, carrying the potential to underpin a new wave of anomaly detection procedures. The work here concentrates on modelling the temporal dependencies governing the time at which events occur. Timestamps are commonly available

with most forms of computer network data and as a result the models described here are widely applicable. Temporal anomaly detection techniques, such as those described in Neil et al. (2013) and Turcotte et al. (2014) have had some success in detecting malicious actors navigating around a network, using much simpler models for the arrivals of events. However, in practice many reported anomalies can be false alarms, where the detectors identify routinely occurring behaviours which have not been captured by the model. This article presents a more flexible modelling procedure, to better capture the characteristics peculiar to computer network traffic. Constructing bespoke and accurate models for normal network behaviour makes it increasingly difficult for a malicious actor to replicate a normal users behaviour without being detected.

The time-based methods considered in this paper are complementary to existing content-based anomaly detectors (see Jin et al. 2003; Singh et al. 2010; Turcotte et al. 2016 for examples). The ultimate aim of any anomaly detection scheme should be to combine the signal of many varying anomaly detection methodologies modelling the many different features available. Additionally, legitimate privacy concerns and the arrival of increasingly ubiquitous encryption together limit the scope that content-based methods can offer to a network analyst, meaning methods such as the

✉ Nicholas A. Heard
n.heard@imperial.ac.uk

Matthew Price-Williams
m.price-williams14@imperial.ac.uk

¹ Department of Mathematics, Imperial College London, London, UK

² Heilbronn Institute for Mathematical Research, University of Bristol, Bristol, UK

present work which rely only upon meta-data have increasing value.

Previous attempts at modelling network traffic have utilised Poisson or Poisson-related models, operating either in continuous time or using time interval binned counts. Lambert et al. (2001) modelled timing patterns in network data using a *dynamic Poisson timing model* with time partitioned into seasonal periods and incorporating an exponentially weighted moving average intensity parameter. Thatte et al. (2008) simply assumed a homogeneous Poisson process to search for attacks through increased activity. Lambert and Liu (2006) proposed a negative binomial model for binned network counts, with control charts for anomaly detection using randomised p values to handle discreteness. To accommodate the heavy-tailed count distributions of network traffic data, Turcotte et al. (2014) also used a negative binomial model to estimate the number of events in a time bin to identify anomalous edges in a network graph.

Whilst these models are computationally convenient, a more realistic modelling approach should seek to capture human user-driven data features such as seasonality and self-exciting behaviour, which are commonly present in computer network traffic data where events can be seen to occur in bursts. These bursts are partly due to users having sessions of activity connecting from a particular client to a particular server, and partly due to the data collection process where a stream of packets are aggregated into grouped summaries of packet flows with some arbitrariness in how the records are divided. This article proposes a general counting process statistical framework aiming to more accurately model the normal traffic generated by each computer within an enterprise network.

For a stream of network traffic, let $Y(t)$ be a counting process recording the number of connections observed by time t , and let $\lambda_Y(t)$ be the conditional intensity of $Y(t)$. Two structural formulations of this intensity will be considered. Conditional on its parameters, the first formulation does not rely on the history of the process but seeks to capture seasonal variation in the arrivals of connections. Supposing behaviours are assumed to be cyclical with season length S , then $\lambda_Y(t)$ will be assumed to have the form

$$\lambda_Y(t) = \mu(t \bmod S), \tag{1}$$

where $\mu(\cdot)$ is a fixed, repeating intensity on $[0, S)$ capturing diurnality and other seasonal fluctuations in the process. The second formulation relaxes this rigid representation of seasonality and instead captures self-exciting behaviour in the process. Letting y_1, y_2, \dots be the increasing sequence of observed event times, the conditional intensity under the self-excitation model will be assumed to depend on the times of the r most recent events, for some $r > 0$, and take the additive form

$$\lambda_Y(t) = \lambda + \sum_{i=1+\max\{0, Y(t)-r\}}^{Y(t)} \omega(t - y_i). \tag{2}$$

The non-negative, non-increasing excitation function $\omega(\cdot)$ controls the increase and subsequent rate of decay in the intensity following an event in the process. Largely for convenience, this is most commonly assumed to be an exponential decay.

1.1 Self-exciting processes

Two extremes of the self-exciting model in (2) are commonly considered in the literature: the Hawkes process (Hawkes 1971), which corresponds to $r = \infty$, and the Wold Process (Wold 1948), where $r = 1$. The majority of work in the literature focuses on the Hawkes process model.

The excitation function $\omega(\cdot)$ is often estimated using a parametric form. In Fox et al. (2016) a Hawkes process with exponential excitation function is used to model the arrivals of emails in a large enterprise network, motivating the use of self-excitation models for computer network data. Hawkes process models with exponential decay excitation have also previously been applied in modelling network structural dependencies (Etesami et al. 2016), corporation default clustering (Azizpour et al. 2017) and earthquake occurrences (Ogata 1988). The work of Linderman and Adams (2014) uses a multivariate Hawkes process with a latent random graph structure to estimate the excitation effect between pairs of processes, capturing the triggering effects generated by trades on the stock market.

Whilst parametric models are convenient, they assume highly constrained forms for the excitation function; there are many scenarios where the parametric model assumptions fail to describe the actual excitation effect. The work of Xu et al. (2016) aims to provide a nonparametric estimation procedure for the excitation functions of a multivariate Hawkes process, with estimation achieved by solving a set of Euler–Lagrange equations. In practice, the method described in Xu et al. (2016) requires more than 10^4 arrivals in order to obtain good results (Yang et al. 2018) and obtaining this quantity of data is often infeasible for training a computer network model.

The present article focuses on the simpler univariate self-exciting Wold process model and in particular proposes a more flexible, nonparametric representation of the excitation function in (2) for modelling computer network traffic. This excitation effect assumes $\omega(\cdot)$ to be a monotonically decreasing step function with an unbounded number of changepoints locating the steps. Allowing an unbounded number of changepoints admits arbitrarily good approximation of the underlying, possibly smooth excitation function as the number of training observations is increased.

For clarity, it should be noted that changepoint methods are used in this article as a flexible modelling tool for approximating a possibly smooth but unknown function; this is an approach commonly exploited by classification and regression trees (Breiman 1984). Such methods benefit from operating on a theoretically non-increasing time domain, leading to so called *infill* asymptotics where the underlying function can be consistently estimated. This is in contrast to the application of changepoint analysis in anomaly detection, outside the scope of this article, where a time-increasing domain typically leads to no such consistency results.

The structure of the remainder of the article is as follows: Sect. 2 describes the motivating network traffic data, obtained from the computer networks of Imperial College London and Los Alamos National Laboratory. Section 3 introduces a piecewise constant model for capturing seasonality in computer network traffic and Sect. 4 examines different models for self-exciting behaviour. Section 5 describes how changepoint estimation algorithms can be used to estimate the parameters of both the seasonal and self-exciting models. The validity of the nonparametric excitation model is assessed through a simulation study in Sect. 6 and the performance of all approaches to modelling computer network traffic are compared on real data in Sects. 7 and 8.

2 Computer network data

Two sources of computer network traffic data are considered: Network flow (“NetFlow”) data from the Imperial College London (ICL) computer network, and authentication logs obtained from the enterprise network of Los Alamos National Laboratory (LANL) (Turcotte et al. 2017). The latter have been suitably anonymised and are publicly available from <https://csr.lanl.gov/data/2017.html>.

NetFlow records are a high level aggregation of the packets sent between two IP addresses over the same ports, under the same protocol in a short space of time, and summarise what might be interpreted as a single communication between the two addresses. The authentication logs specify the times at which users performed authentication-related actions on computers within their internal network. An event type is also recorded, such as network log-on, interactive log-on, workstation screen lock, and so on. In Turcotte et al. (2017) and Price-Williams et al. (2017) it was noted that certain authentication event types exhibit strong periodic patterns, indicating automated network activity. For this analysis, we only consider event types that are considered “interactive” and require a user being present at a computer.

Figure 1 plots examples of each data source. In the left image (a), NetFlow event start times (recorded in milliseconds) are plotted for connections from one ICL IP address to a particular internet sever over a 90-day period. The right

image (b) plots the “interactive log-on” events (recorded in seconds) over a different 90-day period for a random user from the LANL network. It is apparent that the event times in both examples exhibit seasonal variations, with all events occurring within the typical hours of a working day, and clear two day breaks that correspond to weekends. More interestingly, the events are seen to occur in bursts; this is particularly apparent in the NetFlow data.

Note that the Imperial NetFlow data are considered confidential and have not been made available for release, but readers interested in reproducing similar analyses to the NetFlow results described in Sect. 7 might consider the LANL network NetFlow data which are also available from the URL above, although these do not contain external connections to the wider internet and timestamps are rounded to the second.

3 Modelling seasonality

Seasonal patterns often appear in computer network traffic since human users are more likely to be active on weekdays during the day time than they are at night time or at the weekend. In the counting process conditional intensity model (1), these seasonal variations are represented by the seasonal intensity $\mu(\cdot)$ on a specified season $[0, S)$. Here it will be assumed $S = 604,800$ seconds, corresponding to a seasonal intensity which repeats each week.

Within a single week there will be daily fluctuations in activity, with weekend days in particular likely to witness fewer network events than weekdays. However, besides these fluctuations in volume, it will be further assumed that the activity patterns for each day will have the same shape. Let $S' = 86,400$ s, the length of one day. Then formally, we assume

$$\mu(t) = \mu'(t \bmod S') \tilde{\mu}_{\lfloor t/S' \rfloor \bmod 7}$$

for $t \in [0, S)$, where $\mu'(\cdot)$ is a circular probability density function on $[0, S')$, $\lfloor x \rfloor$ is the integer part of x , and $\tilde{\mu}_0, \dots, \tilde{\mu}_6$ are non-negative daily multipliers such that the seasonal conditional intensity (1) on the i th day of the week at the time of day s seconds is $\mu'(s) \tilde{\mu}_i$.

Given some historical event data, the parameters $\tilde{\mu}_0, \dots, \tilde{\mu}_6$ are trivially estimated by the average number of events occurring on each week day, and so it simply remains to estimate the density function $\mu'(s)$ of event arrivals throughout the day. For computational tractability and simplicity, $\mu'(s)$ will be assumed to be piecewise constant with an unknown number of changepoints m , denoted $\sigma_1, \dots, \sigma_m$, ordered such that $0 \equiv \sigma_0 < \sigma_1 < \dots < \sigma_m < \sigma_{m+1} \equiv S'$. Let $\mu'_1, \dots, \mu'_{m+1}$ be the corresponding densities in each changepoint segment, such that

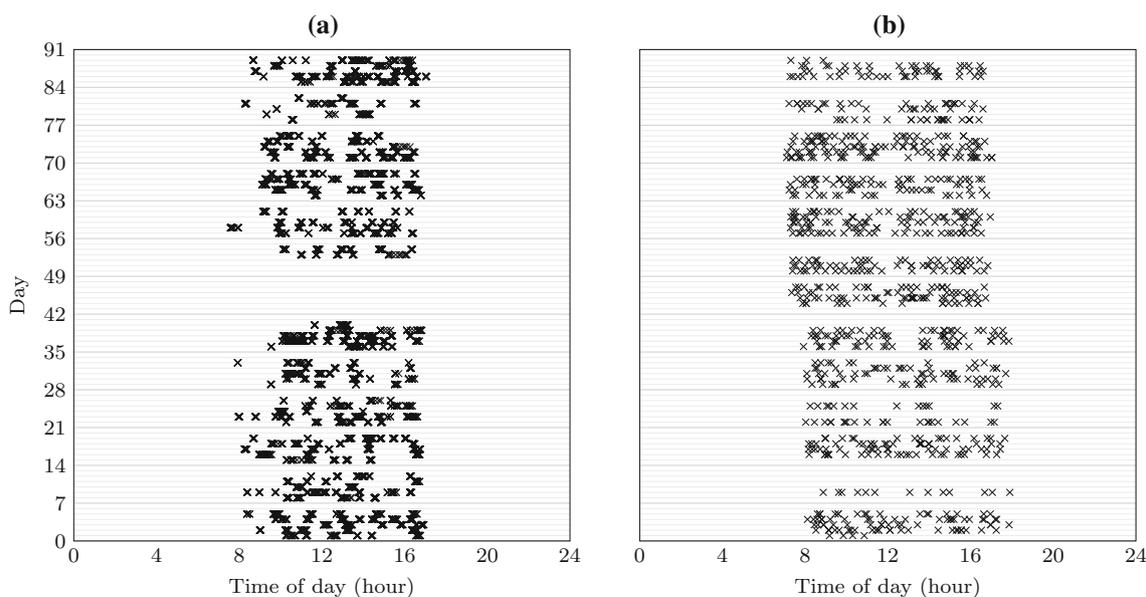


Fig. 1 Examples of user-driven computer network event data: **a** Start times of all connections between one ICL IP address and a particular internet server. **b** Times of all interactive log-on events made by User 715003 from the LANL computer network

$$\mu'(s) = \sum_{j=1}^{m+1} \mu'_j \mathbb{I}_{[\sigma_{j-1}, \sigma_j)}(s), \tag{3}$$

and $\sum_{j=1}^{m+1} \mu'_j (\sigma_j - \sigma_{j-1}) = 1$. Estimation of (3) is deferred until Sect. 5.

4 Modelling self-exciting behaviour

Next consider an alternative formulation of the conditional intensity function (2) modelling self-exciting behaviour. Note that unlike Sect. 3, there are no seasonal assumptions in this model. The conditional intensity under the Hawkes model where $r = \infty$ has an additive contribution from every prior event in the process, whereas the Wold process where $r = 1$ only considers the time elapsed since the most recent event, with simplified conditional intensity

$$\lambda_Y(t) = \lambda + \omega(t - y_{Y(t)}). \tag{4}$$

The Hawkes and Wold variants of (2) will be compared for samples of computer network traffic, first using the standard exponential excitation function

$$\omega(u) = \gamma \exp(-\beta u), \tag{5}$$

where $\gamma, \beta > 0$. Although less common, the exponential excitation function (5) can also be used in the conditional intensity (2) to model self-exciting behaviour for any finite $r > 1$.

Additionally, the simpler Wold model will be compared under an alternative, nonparametric excitation function spec-

ified as a non-increasing step function with an unbounded number of changepoints $\ell \geq 0$. Denoting the changepoints $0 \equiv \tau_0 < \tau_1 < \dots < \tau_\ell$ and the corresponding step heights as $\lambda_1 > \dots > \lambda_\ell$, the proposed excitation function is

$$\omega(u) = \sum_{j=1}^{\ell} \lambda_j \mathbb{I}_{[\tau_{j-1}, \tau_j)}(u). \tag{6}$$

Under the Wold model, the excitation function (6) implies a conditional intensity

$$\lambda_Y(t) = \lambda + \sum_{j=1}^{\ell} \lambda_j \mathbb{I}_{[\tau_{j-1}, \tau_j)}(t - y_{Y(t)}). \tag{7}$$

Illustrations of the Hawkes and Wold process models with exponential decay excitation and the nonparametric Wold process model are shown in Fig. 2.

4.1 Parameter estimation

Given event times y_1, y_2, \dots, y_n from the counting process $Y(t)$ observed on $[0, T]$, from Daley and Vere-Jones (2007) the likelihood function for a given conditional intensity function $\lambda_Y(t)$ is

$$\exp\left(-\int_0^T \lambda_Y(u) du\right) \prod_{i=1}^n \lambda_Y(y_i).$$

Numerical maximum likelihood estimation of the conditional intensity parameters is straightforward for the models with

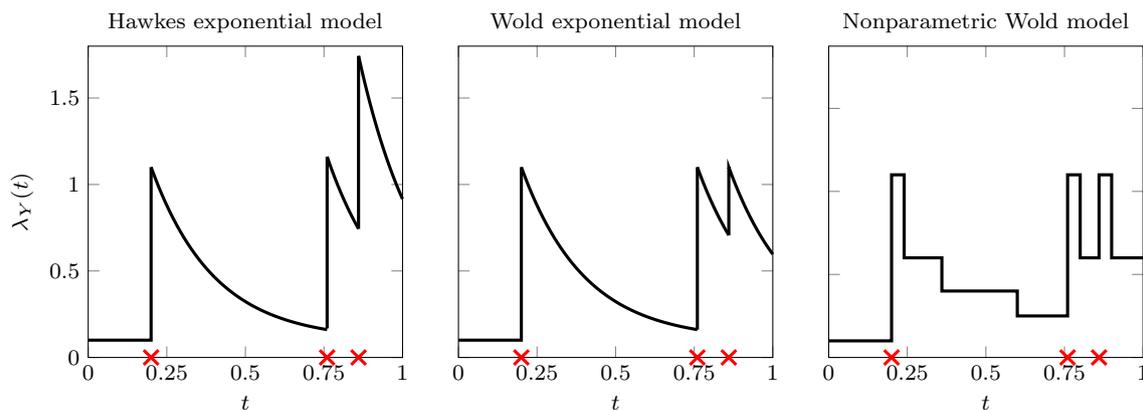


Fig. 2 Illustrative examples of realised conditional intensity functions for a synthetic point process with three events, indicated by crosses

an exponential excitation function (5), which have just three unknown parameters. More details on the estimation procedure are provided in Ozaki (1979), Laub et al. (2015).

In contrast, the nonparametric excitation function (6) has an unbounded number of changepoint and intensity parameters to optimise. Note that the excitation function (6) is only considered for use with the Wold process model (4), since in this case the parameters can be estimated using standard changepoint estimation methodology applied to the process inter-arrival times, as described in the next section.

5 Changepoint estimation for the conditional intensity function

Estimation of the changepoints for the seasonal density (3) or the Wold process with nonparametric excitation function (7) can both be performed very efficiently using the pruned exact linear time (PELT) algorithm of Killick et al. (2012). For an observed sequence of random variables $x_{1:n}$, PELT finds changepoints $\tau_{1:\ell}$, $0 \equiv \tau_0 < \tau_1 < \dots < \tau_\ell < \tau_{\ell+1} \equiv n$, which minimise the Bayesian information criterion (Schwarz 1978),

$$\sum_{j=1}^{\ell+1} \{-2 \log \mathcal{L}(x_{\tau_{j-1}+1:\tau_j}) + \alpha \log n\}, \tag{8}$$

where \mathcal{L} is an estimated likelihood function and $\alpha > 0$ notionally represents the number of additional free parameters introduced to the model by adding a changepoint. In the case where $x_{1:n}$ are the event times from an inhomogeneous Poisson process with piecewise constant intensity,

$$\mathcal{L}(x_{\tau_{j-1}+1:\tau_j}) = \left(\frac{\tau_j - \tau_{j-1}}{x_{\tau_j} - x_{\tau_{j-1}}} \right)^{(\tau_j - \tau_{j-1})} e^{-(x_{\tau_j} - x_{\tau_{j-1}})} \tag{9}$$

and $\alpha = 2$, since introducing a new changepoint adds two parameters: the location of the changepoint, and the intensity within the new segment.

5.1 Estimating the seasonal density function

Since the seasonal density (3) in Sect. 3 is a circular density, to estimate $\mu'(\cdot)$ historical event times from $Y(t)$ are mapped onto $[0, S')$ with respect to a variable origin shift s_0 . Estimation then proceeds applying PELT to changepoint modelling of the order statistics of the wrapped event times $\{(y_i - s_0) \bmod S'\}$ as an inhomogeneous Poisson process (cf. the likelihood (9)) on $[0, S')$ with piecewise constant intensity which will be proportional to the final fitted density. Finally the cost function (8) is minimised over all possibly optimal origin shifts $s_0 \in \{0, y_1, \dots, y_n\}$.

5.2 Estimating the nonparametric excitation function

Novel methodology is required for estimating the conditional intensity function of the nonparametric Wold process model defined in (7), consisting of three steps. First, a time-rescaling transformation is applied to the inter-arrival times to form an inhomogeneous Poisson process. Second, the maximum likelihood estimate (MLE) changepoints and intensities are identified subject to the constraint that the estimated and intensity levels are non-increasing (Boswell 1966). Third, to guard against over-fitting, PELT is then applied to the same changepoint problem with candidate changepoints constrained to be selected from the MLE changepoints found in the previous step, guaranteeing monotonicity in the intensity as well as a parsimonious model. These steps are now explained in more detail.

Step 1 Let $y_1, \dots, y_n \in Y(t)$ be event times from a Wold process model with conditional intensity

$$\lambda_Y(t) = \lambda + \sum_{j=1}^{\ell} \lambda_j \mathbb{I}_{(\tau_{j-1}, \tau_j]}(t - y_Y(t)),$$

Let $d_i = y_{i+1} - y_i$ be the waiting time between successive events in Y . Then d_1, \dots, d_{n-1} are independent identically distributed variables with hazard function

$$h(t) = \lambda + \sum_{j=1}^{\ell} \lambda_j \mathbb{I}_{(\tau_{j-1}, \tau_j]}(t). \tag{10}$$

For maximum likelihood estimation, the changepoint locations can be restricted to the observed waiting times d_1, \dots, d_{n-1} . If $d_{(1)}, \dots, d_{(n-1)}$ are the corresponding order statistics, then (10) simplifies to

$$h(t) = \lambda + \sum_{j=1}^{\ell} \lambda_j \mathbb{I}_{(d_{(\tau_j^*)}, d_{(\tau_{j-1}^*)}]}(t), \tag{11}$$

where $\tau_j^* \in \{1, \dots, n-1\}$, $j = 1, \dots, \ell$. To estimate the parameters of (11) and consequently the conditional intensity (7), firstly set $d_{(0)} \equiv 0$. Then for $i = 1, \dots, n-1$, the random variable

$$\delta_i = (n-i)(d_{(i)} - d_{(i-1)})$$

is a draw from an exponential distribution with rate $\lambda + \lambda_{\min_j \{\tau_j^* > i-1\}}$. Defining $\Delta_0 \equiv 0$, $\Delta_i = \Delta_{i-1} + \delta_i$, then $\Delta_1, \Delta_2, \dots$ can be seen as the event times of an inhomogeneous Poisson process with piecewise constant non-increasing conditional intensity function

$$\lambda + \sum_{j=1}^{\ell} \lambda_j \mathbb{I}_{(\Delta_{\tau_{j-1}^*}, \Delta_{\tau_j^*}]}(u). \tag{12}$$

Step 2 It is well known (see, for example, Boswell 1966) that maximum likelihood estimates for the changepoints of (12), subject to the constraints $\lambda_1 > \lambda_2 > \dots > \lambda_{\ell}$, can be calculated recursively as

$$\hat{\tau}_j = \arg \max_{i \in \hat{\tau}_{j-1}+1, \dots, n-1} \left(\frac{i - \hat{\tau}_{j-1}}{\Delta_i - \Delta_{\hat{\tau}_{j-1}}} \right), \tag{13}$$

terminating when $\hat{\tau}_j = n-1$. Let $\hat{\ell}$ be the resulting estimate of the number of changepoints. Then the corresponding estimated intensity levels are

$$\hat{\lambda} = \frac{(n-1) - \hat{\tau}_{\hat{\ell}}}{\Delta_{n-1} - \Delta_{\hat{\tau}_{\hat{\ell}}}}$$

$$\hat{\lambda}_j = \frac{\hat{\tau}_j - \hat{\tau}_{j-1}}{\Delta_{\hat{\tau}_j} - \Delta_{\hat{\tau}_{j-1}}} - \hat{\lambda}, \quad j = 1, \dots, \hat{\ell}.$$

Step 3 Even though the maximum likelihood estimation in the previous step is constrained by monotonicity, there is still a possibility of overfitting the data by overestimating $\hat{\ell}$. To guard against this, the PELT algorithm with likelihood function (9) is finally applied to the event sequence $\Delta_{1:n-1}$, where the changepoints are restricted to be a subsequence of the values $\hat{\tau}_{1:\hat{\ell}}$ from (13). By restricting to these candidate changepoints, a non-increasing intensity function is still guaranteed.

6 Simulation study for the Wold process model with nonparametric excitation

A simulation study is performed to empirically investigate the estimation error for the inference procedure of Sect. 5 for the Wold process with nonparametric excitation function (7). For each $n \in \{10, 100, 1000, 10000\}$, $2n$ event times y_1, \dots, y_{2n} are drawn from the Wold process model (7) with parameters $\lambda = 0.001$, $\ell = 5$, $\tau_{1:\ell} = (0.5, 2.0, 4.0, 20.0, 100.0)$, $\lambda_{1:\ell} = (1.0, 0.5, 0.1, 0.05, 0.01)$. A non-trivial number of changepoints ($\ell = 5$) are chosen to represent the potential complexity of self-excitation in computer network traffic. For each simulated sequence of event times, the parameters of the conditional intensity function are estimated from the first n event times using the procedure described in Sect. 5 and then the goodness-of-fit is examined on the last n event times.

For estimating goodness of fit of the conditional intensity function (7), define the compensator function $\Lambda(t) = \int_{s=0}^t \lambda_Y(s) ds$. Under a null hypothesis that the estimated intensity is correct, the time rescaling theorem (Brown et al. 2002) states that $\Lambda(y_1), \Lambda(y_2), \dots$ must be the event times of a homogeneous Poisson process with unit rate. Calculating a lower-tail p value of the corresponding waiting times,

$$p_i = 1 - \exp[-\{\Lambda(y_i) - \Lambda(y_{i-1})\}], \tag{14}$$

it follows that $p_i \sim \text{Uniform}(0, 1)$ under that null hypothesis. Alternatively, consistently small (large) p values suggest smaller (larger) than expected waiting times between events.

After estimating the conditional intensity (7) from the first n event times, the p values (14) are calculated for the final n event times a_{n+1}, \dots, a_{2n} for checking model fit: Q-Q plots of the p values for an example simulation are displayed in Fig. 3 for the different values of n . For a moderate sample size of $n = 1000$ the distribution of the p values is very close to uniform, and even for $n = 100$ the distributional approximation appears reasonable.

To ascertain consistency of the estimated parameters, the analysis was repeated for 1000 simulated event time

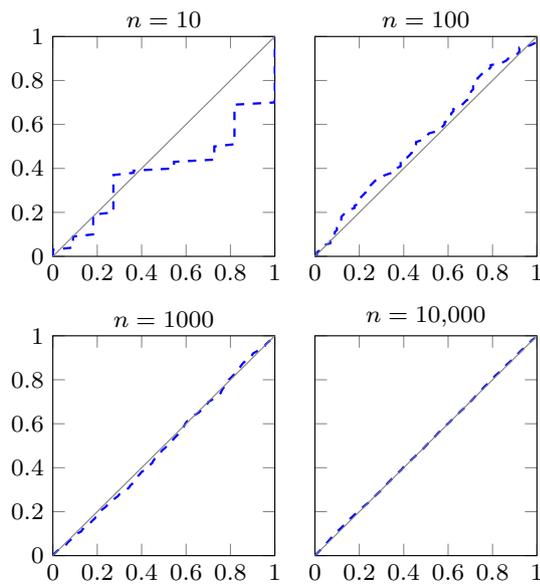


Fig. 3 Q–Q plots of the p values obtained from the final n event times simulated from the generative model of the nonparametric excitation Wold process model (7)

Table 1 The mean squared error between the real and estimated intensity at 100 evenly spaced times, calculated over 1000 simulated event sequences

n	10	100	1000	10000
MSE	0.15	2.0×10^{-3}	2.0×10^{-4}	5.8×10^{-5}

sequences to calculate the mean square error between the true and estimated intensities over 100 evenly spaced time points, shown in Table 1. Consistent estimation appears to be taking place with the errors shrinking rapidly with n . Other simulations were completed with alternative parameters for the intensity, yielding similar results.

7 Modelling NetFlow event times

To investigate which of the model specifications for the conditional intensity function (1) and (2) best represents normal computer network traffic behaviour, the predictive performances of the rival models are compared on 90 days of network traffic (NetFlow) data from the Imperial College London (ICL) computer network, described in Sect. 2. Focusing on the internet server shown to characterise human-like behaviour in Fig. 1a, 26 client IP addresses were observed to connect to this server at least 200 times in both the first 28 and the last 62 days of the data collection. All client IP addresses displayed similar human behaviour with a clear diurnal connectivity pattern.

For each client IP address, the parameters of the conditional intensity function for both models are estimated using the first four weeks of the NetFlow data, and the goodness-of-fit is compared using the distribution of p values (14) calculated for the remaining 62 days using the procedure described in Sect. 6. Specifically, a Kolmogorov–Smirnov (KS) test (Massey Jr 1951) is used to assess the goodness-of-fit by measuring the supremum absolute difference between the cumulative distribution functions of the observed p values and the Uniform(0, 1) distribution; smaller values of this statistic correspond to better model fit.

As a baseline for comparison, a homogeneous Poisson process (HPP) model, suggested in Thatte et al. (2008), is fitted to the event data for each of the 26 clients; the maximum likelihood estimate for the HPP intensity parameter is simply the mean number of events observed per unit of time. This baseline HPP model is then compared to the model for seasonal behaviour from Sect. 3 and the Hawkes and Wold process models for self-exciting behaviour described in Sect. 4. Additionally, these models are compared to the generalised self-exciting model (2) with exponential excitation function (5) and summing excitation effects over the most recent r event times for $r \in \{2, 5, 10\}$.

Figure 4 displays box plots of the KS test statistics for each model across the client IP addresses. The Wold model with nonparametric excitation function outperforms all other methods for 24 of the 26 client IP addresses considered. The self-exciting models which are estimated using the most recent five or ten events are omitted from this figure and the subsequent analyses because the resulting box plots are visually indistinguishable from the Hawkes model ($r = \infty$).

It is interesting to note that the seasonal model actually performs worse than the simple homogeneous Poisson model for some client IP addresses. This is because these clients are not regularly active on the server in question, and the time of day that they access the server varies; the seasonal behaviour in the first 28 days is therefore not representative of the time of day activity in the remaining 62 days. This highlights the inflexibility of a seasonal modelling approach, and partly justifies the absence in this comparison of hybrid models including both seasonal and self-excitation effects (see Fox et al. 2016); furthermore, seasonal and self-excitation effects cannot be jointly estimated using the same efficient procedure detailed in Sect. 5.

For a visual illustration, Q–Q plots of the p values for the traffic sequence in Fig. 1a are presented in Fig. 5a. The seasonal model can be seen to provide only a limited improvement over the assumption of a constant intensity; the shape of both of these Q–Q curves imply p values which are too small, meaning the majority of the events arrive more quickly than expected. In contrast, any of the models for capturing self-excitation provide a significant jump in performance. Furthermore, the Wold model with nonparametric

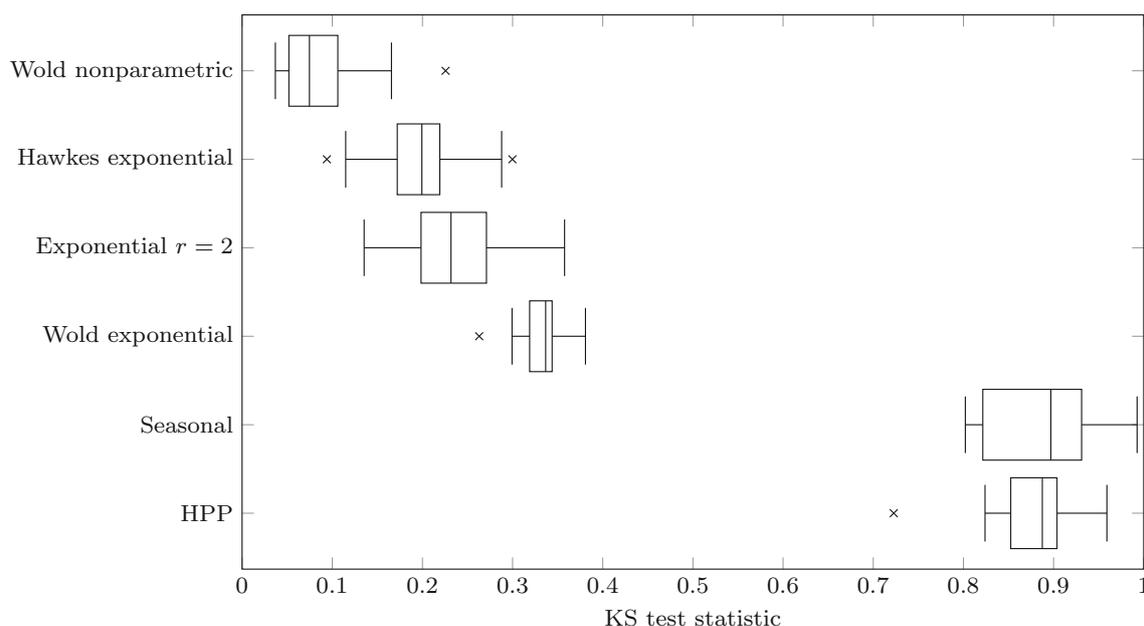


Fig. 4 Box plots of the KS test statistics from modelling NetFlow connections between 26 network clients and an internet server, under different models for the conditional intensity function

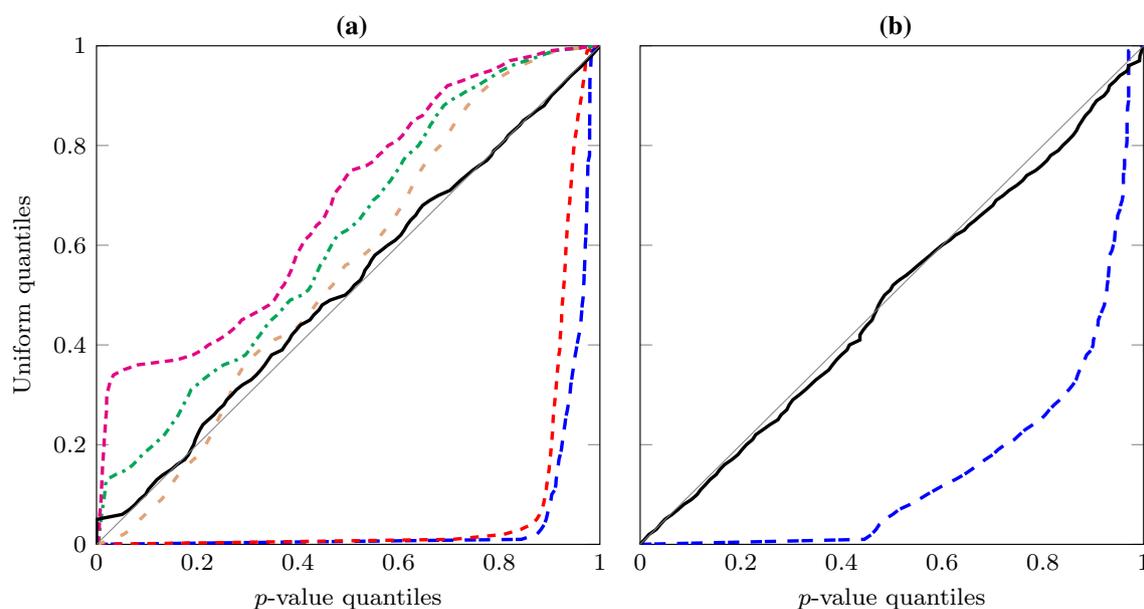


Fig. 5 Q–Q plots of the p values (14) for different conditional intensity models: HPP (---); seasonally adjusted (-.-.); a Hawkes model with exponential excitation function (-.-.); a Wold model with exponential (---) or nonparametric (—) excitation function or an exponential

self-exciting model using the two most recent events (-.-.-). **a** Connections from a client IP address in the ICL computer network to an internet server. **b** Log-on event times from user 715003 in the LANL computer network

excitation outperforms all models using an exponential excitation function, with the unbounded number of parameters offering greater flexibility for capturing any underlying self-excitation.

Figure 6 shows the fitted hazard function (10) for waiting times from the traffic sequence shown in Fig. 1a. There are

14 fitted changepoints in this example, with the last changepoint occurring after about two and a half hours, after which there is no remaining self-excitation effect. The distribution of the number of fitted changepoints for the 26 clients is shown in Fig. 7 as a scatter plot against the number of training data points occurring in the first four weeks; as usual

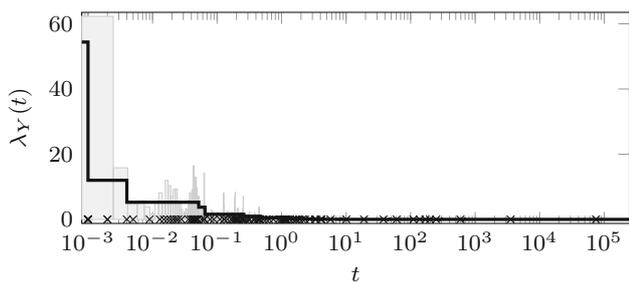


Fig. 6 Example fitted hazard function (10) under the nonparametric Wold model, shown on a log time scale. The histogram shows the distribution of waiting times between events

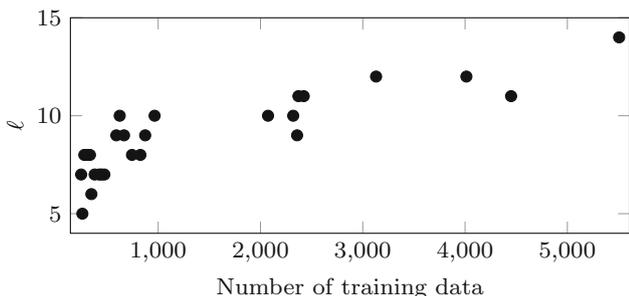


Fig. 7 The number of changepoints ℓ minimising the cost function (8) under the nonparametric Wold model (7), plotted against the number of event times available to fit the model

with nonparametric models, the number of fitted parameters (changepoints) can be seen to generally increase with the number of data points, as the true (possibly smooth) underlying self-excitation function is captured increasingly well.

7.1 Anomaly detection

The principal aim has been to demonstrate how well inter-arrival times can be modelled under normal conditions. However, to give a very simple demonstration of the potential for temporal modelling techniques to be applied in cybersecurity anomaly detection, a brief example is now presented.

In the absence of any ground truth data from an actual cyber attack within the Imperial NetFlow collection data, a hypothetical attack is constructed by fusing the NetFlow event times from two different network edges, imitating a sudden change in connectivity patterns. The parameters of the Wold process model with nonparametric intensity are initially estimated using 28 days of event times from an IP address connecting through a TCP/IP port to a particular internet server using SSL secure encryption. For testing, event times from the final 62 days are collected from TCP/IP connections between the same IP address and a different HTTP server from the same web domain, using unencrypted communication; this change in both the destination IP address and the encryption used should lead to a

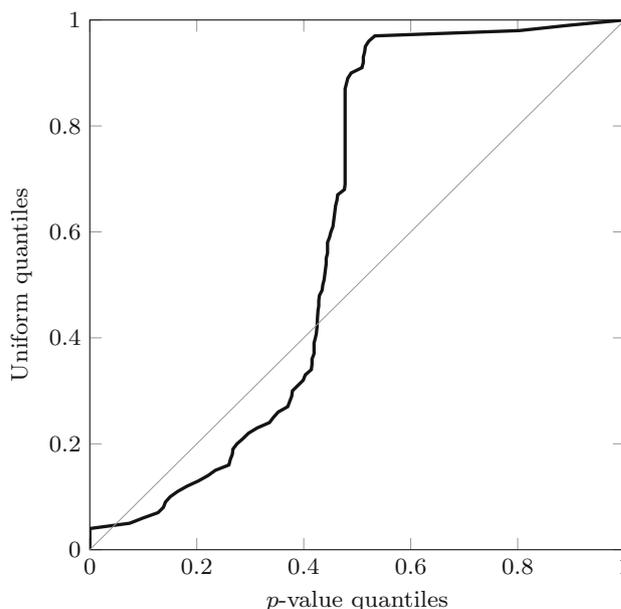


Fig. 8 Q-Q plot of the nonparametric Wold model p values (14) calculated for anomaly detection, when training parameters for 28 days on one edge and testing the resulting predictions for 62 days on another edge

change in connectivity patterns. To analyse the performance from conducting anomaly detection using the fitted Wold process model, p values are calculated for the inter-arrival times from the test data of HTTP NetFlow timestamps.

The corresponding Q-Q plot from the test data is shown in Fig. 8. The KS statistic is 0.4416 which is significantly larger than any of the KS statistic values which were calculated when predicting the same internet server in both the training and test periods for model checking (see Fig. 4).

8 Modelling discrete authentication log times

8.1 Handling discrete event times

For some computer network architectures and data sources, event times may be recorded more crudely than the NetFlow data in Sect. 7. In particular, the authentication logs from the LANL enterprise network introduced in Sect. 2 have event times recorded only to the nearest second. Particularly when the event times occur in bursts, this can lead to multiple events appearing to occur simultaneously, in which case a Poisson process model is no longer appropriate. One solution could be to filter all events that occur in the same second and treat them as a single, combined event in any further analysis. A problem with this solution, particularly in anomaly detection, is that a sequence of events that occur in bursts may actually represent anomalous behaviour and binning the events within

one second may weaken or remove the anomalous signal. Instead, the approach considered here is to explicitly model the arrivals as a discrete time process.

For brevity, in this section attention is restricted to the nonparametric Wold model (7) which performed best in Sect. 7 and also performs best with these data. This model is particularly straightforward to convert into discrete time.

In continuous time, under the nonparametric Wold model the conditional intensity function (7) corresponds to piecewise exponentially distributed waiting times between events. To recast this model in discrete time, first the changepoints $\tau_{1:\ell}$ are constrained to be discrete and integer-valued; second, since the geometric distribution is the discrete analogue of the exponential distribution, it will be assumed that the waiting times are now piecewise geometrically distributed: Upon entering the discrete time segment $\{\tau_{j-1}, \dots, \tau_j - 1\}$ without having observed an event, the next event may occur at each time point in the segment with constant hazard $\lambda + \lambda_j$. This formulation retains the functional form for the hazard function (10), but this is now interpreted as a discrete hazard which allows an unbounded number of events to occur at each time point, since there is positive mass (hazard) at time zero.

For performing inference, the methodology described in Sect. 5 can be used with the likelihood function (9) adapted to the revised geometric model: For the event times $x_{1:n}$, the corresponding geometric likelihood for PELT is

$$\mathcal{L}(x_{\tau_{j-1}+1:\tau_j}) = \left(\frac{\tau_j - \tau_{j-1}}{x_{\tau_j} - x_{\tau_{j-1}}} \right)^{\tau_j - \tau_{j-1}} \times \left(1 - \frac{\tau_j - \tau_{j-1}}{x_{\tau_j} - x_{\tau_{j-1}}} \right)^{x_{\tau_j} - x_{\tau_{j-1}}}. \tag{15}$$

To see the asymptotic equivalence of the continuous and discrete time models, note that the geometric likelihood (15) converges to (9) as the segment length $x_{\tau_j} - x_{\tau_{j-1}} \rightarrow \infty$.

8.2 Analysis of user authentication event data

Mirroring the comparisons of Sect. 7, the performance of the discrete time nonparametric excitation Wold process is compared with a homogeneous process, in this case assuming a discrete hazard function which is constant over time. Also following Sect. 7, the LANL authentication event data are also divided into four weeks of training data for fitting the intensity models and testing is carried out on the remaining 62 days of data.

To examine predictive performance of the models, lower tail p values are once again calculated for each of the events during the testing period, for all 3119 users in the LANL network for whom there were at least 200 events in both the training and test periods of data collection. Under the discrete

time geometric formulation, the analogous equation to (14) for the lower-tail p value is

$$p_i = 1 - \prod_{t=y_i-1}^{y_i} \{1 - \lambda_Y(t)\}. \tag{16}$$

For assessing model fit, it must be noted that the p values (16) are discrete, and therefore stochastically larger than Uniform(0, 1) under the assumed model. To ease this difficulty, randomised p values (Habiger and Pena 2011) are generated uniformly between the p value (16) for y_i and the corresponding value \tilde{p}_i which would be obtained from observing $y_i - 1$,

$$\tilde{p}_i = 1 - \prod_{t=y_i-1}^{y_i-1} \{1 - \lambda_Y(t)\}.$$

It is easily verified that if the discrete intensity model were true, a random p value drawn from Uniform(\tilde{p}_i, p_i) would be marginally Uniform(0, 1). (Note that when monitoring a traffic sequence in real time for anomaly detection purposes, it may be more appropriate to use the mid p value (Rubin-Delanchy and Heard 2015) to remove randomness in the decision making process without being overly conservative.)

For all 3119 users, a Kolmogorov–Smirnov test (cf. Sect. 7) is conducted for the random p values generated from the predictions arising from each fitted intensity model. The resulting distributions of the KS test statistics are shown in Fig. 9. The discrete-time nonparametric excitation Wold model strongly outperforms the homogeneous discrete hazard model, providing a superior predictive fit for 3103 of the 3119 network users analysed.

Figure 10 shows the distribution of the number of fitted changepoints for all users, again presented as a scatter plot against the number of training data points occurring in the first four weeks. There is again some correlation in these quantities, but more importantly far fewer changepoints are fitted in comparison with the NetFlow data in Sect. 7. Two possible explanations for this are the discretisation of the event times making detection of more subtle dependencies difficult, and also the very different, less complex nature of network authentications compared to network flow traffic; recall in Fig. 1, the example NetFlow data showed more pronounced bursts of events than the authentication logs.

To illustrate the quality of predictive fit under the nonparametric Wold model, Fig. 5b shows Q–Q plots of the random p values for the user interactive log-on times shown in Fig. 1b under the rival discrete intensity models. As with the NetFlow data, the nonparametric Wold model shows excellent fit with near uniform p values. The fitted conditional intensity model for this user has a background discrete hazard of

Fig. 9 Box plots of the KS test statistic distributions from modelling user-driven network behaviour, across 3119 users from the LANL computer network, assuming either a time-homogeneous discrete time hazard or the discrete Wold model with nonparametric excitation

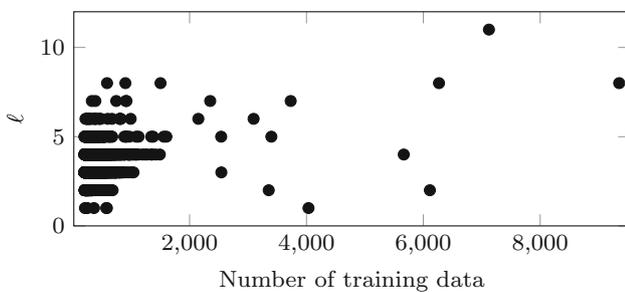
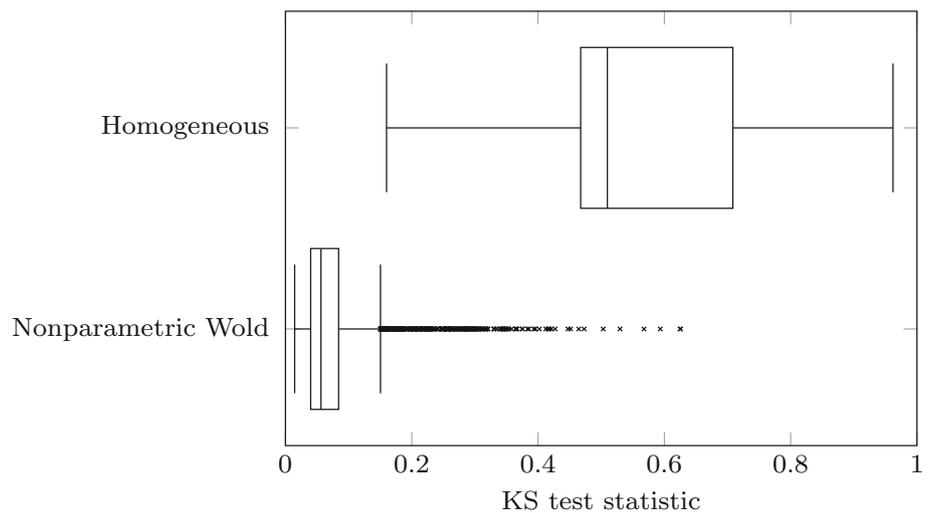


Fig. 10 The number of fitted changepoints under the discrete nonparametric Wold model (7) for each user in the LANL authentication data, against the number of training event times

the order of 10^{-6} of a new authentication event each second, but following an event there is a boost in hazard to around 0.4 probability of another event within the same second; in the next second the hazard drops down to be of order 10^{-3} for another eight seconds and then gradually further decreases.

9 Conclusion

A variety of models have been compared for the conditional intensity of arrivals of computer network traffic events. Building accurate models for these arrival times is an important step towards developing practical statistical analytics for cyber-security which look for deviations from normal network behaviour. Models which try to capture seasonal behavioural patterns have been shown to be too inflexible to capture normal human variation. In contrast, self-exciting intensity models have demonstrated encouraging model fit, with the best performance achieved with a novel nonparametric model which should asymptotically converge to any true underlying excitation function.

10 Supplementary material

Supplementary materials available online at <https://github.com/Matt0312/SToCND> contain python code to implement the methods introduced in the article.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Azizpour, S., Giesecke, K., Schwenkler, G.: Exploring the sources of default clustering. *J. Financ. Econ.* **129**, 154–183 (2017)

Boswell, M.T.: Estimating and testing trend in a stochastic process of Poisson type. *Ann. Math. Stat.* **37**(6), 1564–1573 (1966)

Breiman, L.: *Classification and Regression Trees*. CRC Press, Boca Raton (1984)

Brown, E.N., Barbieri, R., Ventura, V., Kass, R.E., Frank, L.M.: The time-rescaling theorem and its application to neural spike train data analysis. *Neural Comput.* **14**(2), 325–346 (2002)

Daley, D.J., Vere-Jones, D.: *An Introduction to the Theory of Point Processes: volume II: General Theory and Structure*. Springer, Berlin (2007)

Etesami, J., Kiyavash, N., Zhang, K., Singhal, K.: Learning network of multivariate Hawkes processes: A time series approach (2016). [arXiv:1603.04319](https://arxiv.org/abs/1603.04319)

Fox, E.W., Short, M.B., Schoenberg, F.P., Coronges, K.D., Bertozzi, A.L.: Modeling e-mail networks and inferring leadership using self-exciting point processes. *J. Am. Stat. Assoc.* **111**(514), 564–584 (2016)

Habiger, J.D., Pena, E.A.: Randomised p values and nonparametric procedures in multiple testing. *J. Nonparametric Stat.* **23**(3), 583–604 (2011)

Hawkes, A.G.: Point spectra of some mutually exciting point processes. *J. R. Stat. Soc. Ser. B* **33**(3), 438–443 (1971)

- Heard, N., Rubin-Delanchy, P.: Network-wide anomaly detection via the dirichlet process. In: IEEE Big Data Analytics for Cybersecurity Computing (BDAC2016). IEEE (2016)
- Jin, C., Wang, H., Shin, K.G.: Hop-count filtering: an effective defense against spoofed DDoS traffic. In: Proceedings of the 10th ACM Conference on Computer and Communications Security, pp. 30–41. ACM (2003)
- Killick, R., Fearnhead, P., Eckley, I.: Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* **107**(500), 1590–1598 (2012)
- Lambert, D., Liu, C.: Adaptive thresholds. *J. Am. Stat. Assoc.* **101**(473), 78–88 (2006)
- Lambert, D., Pinheiro, J., Sun, D.X.: Estimating millions of dynamic timing patterns in real time. *J. Am. Stat. Assoc.* **96**(453), 316–330 (2001)
- Laub, P.J., Taimre, T., Pollett, P.K.: Hawkes processes (2015). [arXiv:1507.02822](https://arxiv.org/abs/1507.02822)
- Lazarevic, A., Ertöz, L., Kumar, V., Ozgur, A., Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection. In: Proceedings of the 2003 SIAM International Conference on Data Mining, pp. 25–36 (2003)
- Linderman, S., Adams, R.: Discovering latent network structure in point process data. In: International Conference on Machine Learning, pp. 1413–1421 (2014)
- Massey Jr., F.J.: The Kolmogorov–Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**(253), 68–78 (1951)
- Neil, J., Hash, C., Brugh, A., Fisk, M., Storlie, C.B.: Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics* **55**(4), 403–414 (2013)
- Ogata, Y.: Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.* **83**(401), 9–27 (1988)
- Ozaki, T.: Maximum likelihood estimation of Hawkes' self-exciting point processes. *Ann. Inst. Stat. Math.* **31**(1), 145–155 (1979)
- Price-Williams, M., Heard, N.A., Turcotte, M.: Detecting periodic subsequences in cyber security data (2017). [arXiv:1707.00640](https://arxiv.org/abs/1707.00640)
- Rubin-Delanchy, P., Heard, N.A.: On the mid- p value of a test statistic with arbitrary real support (2015). [arXiv:1505.05068](https://arxiv.org/abs/1505.05068)
- Schwarz, G.: Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978)
- Singh, N.K., Tomar, D.S., Roy, B.N.: An approach to understand the end user behavior through log analysis. *Int. J. Comput. Appl.* **5**(11), 27–34 (2010)
- Thatte, G., Mitra, U., Heidemann, J.: Detection of low-rate attacks in computer networks. In: INFOCOM Workshops 2008, IEEE, pp. 321–332. IEEE (2008)
- Turcotte, M., Heard, N.A., Kent, A.D.: Modelling user behaviour in a network using computer event logs. In: Dynamic Networks in Cybersecurity, pp. 67–77. Imperial College Press (2016)
- Turcotte, M., Heard, N.A., Neil, J.: Detecting localised anomalous behaviour in a computer network. In: International Symposium on Intelligent Data Analysis, pp. 321–332. Springer (2014)
- Turcotte, M., Kent, A.D., Hash, C.: Unified host and network data set (2017)
- Wold, H.O.A.: On prediction in stationary time series. *Ann. Math. Stat.* **19**(4), 558–567 (1948)
- Xu, H., Farajtabar, M., Zha, H.: Learning granger causality for hawkes processes. In: International Conference on Machine Learning, pp. 1717–1726 (2016)
- Yang, Y., Etesami, J., He, N., Kiyavash, N.: Nonparametric hawkes processes: online estimation and generalization bounds (2018). [arXiv preprint arXiv:1801.08273](https://arxiv.org/abs/1801.08273)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.