Modelling user-driven network data using Hawkes and Wold processes

Matthew Price-Williams
Dr. Nick Heard

Department of Mathematics, Imperial College London, London, UK

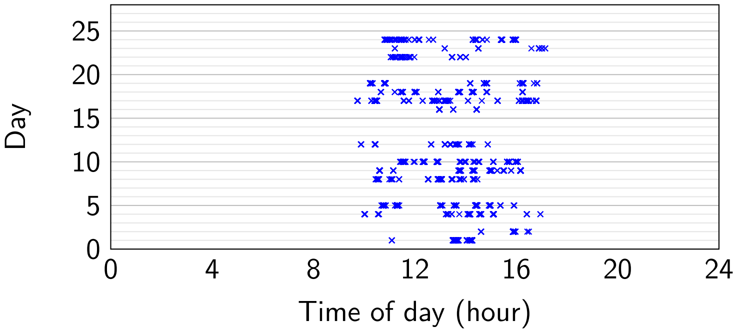
September 2017

Motivation

- **Anomaly detection** involves building probabilistic models of normal computer network behaviour and finding deviations from the model.
- In practice many reported anomalies end up being false.
- One approach to address this issue, which is explored in Grana et al. [2016], is to develop a model of attacker behaviour.
- Additionally we can improve the model of normal network conditions by modelling key features such as **seasonality and self-exciting behaviour.**

Netflow data

- An example of all event times between two IP address is plotted below.
- It is apparent that these data exhibit seasonal patterns and self-exciting behaviour.
- The sequence of event times is modelled as a point process.



Point processes

Let $Y = \{y_1, \dots y_n\}$ be a sequence of points in $[0, T)$, such that $0 \leq y_1 \leq \dots \leq y_n$. Then Y is a **finite point process.**
Let $N_Y(y)$ be the counting process such that

$$N_Y(y) = \#\{y_i \leq y\},$$

One way to characterise a point process Y is by specifying the **conditional intensity function $\lambda^*(t)$.** Specifically

$$f^*(t) = \lim_{h \downarrow 0} \frac{\mathbb{E}[N_Y(t+h) - N_Y(t) \mid \mathbb{H}(t)]}{h},$$

where $\mathbb{H}(t)$ specifies the history of the process Y before time t.

Modelling seasonal behaviour

- Seasonal behaviour is modelled using an **inhomogeneous Poisson process.**
- The conditional intensity function $\mu^*(t)$ is therefore a **step function.**

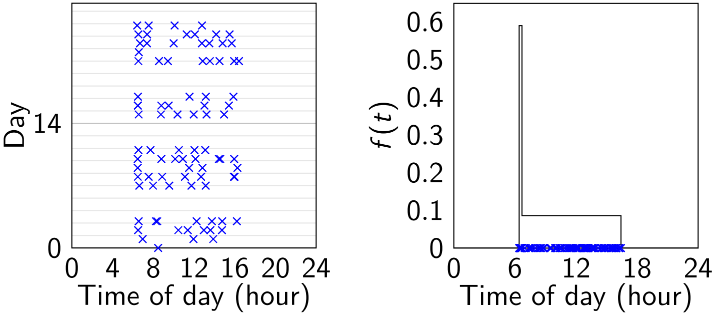


Figure 1: Density plot for the time of day activity for a user X from the LANL computer network.

Self-exciting behaviour

For each realisation y_i in Y, let

$$z_i = \int_0^{y_i} \mu^*(t)dt.$$

Under the Null model of no self-exciting behaviour, Z is a homogeneous Poisson process.
Under the alternative model, **each arrival causes a temporary increase** in the intensity of the point process Z.

The conditional intensity function of Z is defined as $\lambda^*(t), \quad t \in [0, N]$.

4 models for self-exciting behaviour

Two different processes are considered for **self-exciting behaviour.**

A **Hawkes** process is defined by the conditional intensity function

$$\lambda^*(t) = \lambda + \sum_{z_i < t} \omega(t - z_i).$$

The intensity of the point process is dependent on the entire history of the process.

A **Wold** process is defined by the conditional intensity function

$$\lambda^*(t) = \lambda + \omega\left(t - \max_i(z_i \mid z_i < t)\right).$$

The intensity of the point process only depends on the time since the last event.

The excitation function

$\omega(t - z_i)$ is the **excitation function** of the process. Two excitation functions are considered:

1. The intensity of the process decays **exponentially** over time.

$$\omega(t - z_i) = \alpha \exp\left(-\beta(t - z_i)\right) \qquad \alpha, \beta > 0.$$

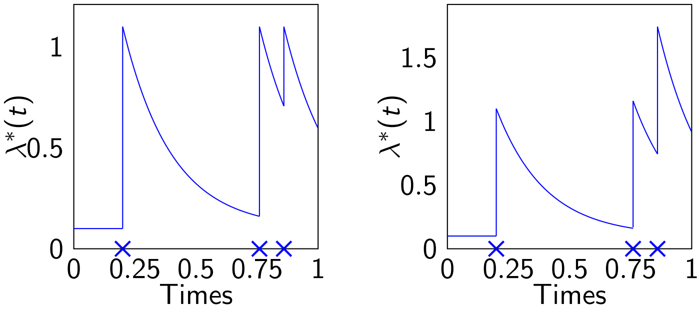
2. The intensity of the process decreases as a **step function**.

$$\omega(t - z_i) = \begin{cases} \lambda_1 & t - z_i \leq \tau_1, \\ \lambda_2 & \tau_2 \geq t - z_i > \tau_1, \\ ... & \\ 0 & t - z_i > \tau_{l-1}, \end{cases} \qquad \lambda_1 > \lambda_2 > \ldots > 0.$$

This model has an unknown number of parameters.

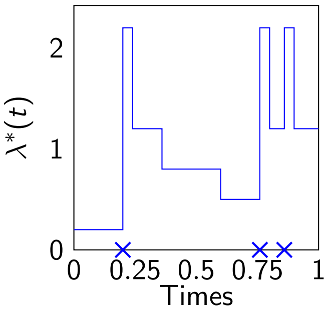
Example Intensity plot for the exponential function model

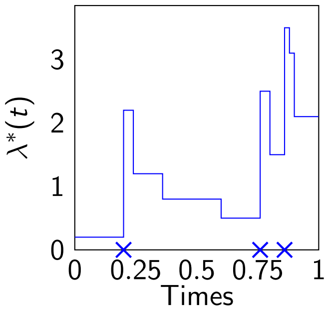
An example of the conditional intensity function is plotted for the Hawkes and Wold models with exponential excitation.



Example Intensity plot for the step function model

The conditional intensity functions are also plotted for the Hawkes and Wold models with a step excitation function.





The number and value of the parameters are estimated using the data.

Parameter estimation

- The parameters of the Hawkes and Wold exponential models are estimated **numerically** using their **maximum likelihood estimates**.
- The parameters of the Wold step function model can be be estimated using a **changepoint detection methodology** to minimise the BIC.
- The Hawkes step function model is simplified and estimated for a fixed number of parameters using the Nelder-mead algorithm.

Assessing performance

- To asses the performance of the model, consider the **waiting times** between two consecutive events, $d_i = y_{i+1} - y_i$.
- For each model let $q_i = \mathbb{P}(d > d_i)$ be the i^{th} **upper tail p-value** specifying the probability that we would have seen a waiting time of greater that d_i under the null model.
- If the model is accurate, $q_i \overset{iid}{\sim} U[0, 1]$.

Results for the seasonal model

The model incorporating seasonality is compared to a homogeneous Poisson process model. Neither model is an appropriate fit for the data.

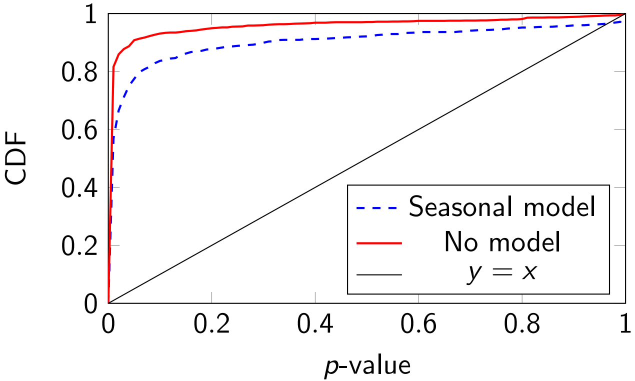


Figure 2: p-values from the inter-arrival times on one edge in a computer network.

Results for the exponential model

The Hawkes and Wold exponential models for self-exciting behaviour are compared below.

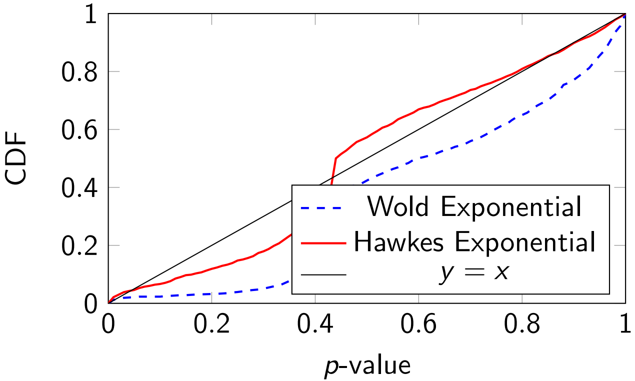


Figure 3: p-values from the inter-arrival times on one edge in a computer network.

Results for the step function models

Additionally the Hawkes and Wold step function models are compared on the data.

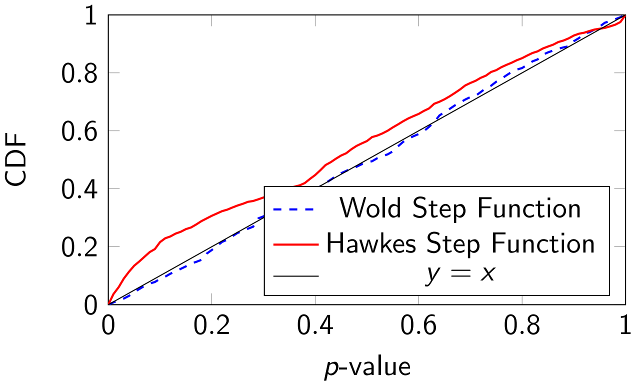


Figure 4: p-values from the inter-arrival times on one edge in a computer network.

Future work and Mutually exciting processes

- For future work, we are interested in **jointly modelling** separate correlated edges in a computer network.
- For instance a workstation unlock event may cause an increase in the intensity of logon events.
- But a workstation lock event may inhibit a logoff event.
- Separate point processes can be modelled jointly using **multivariate** Hawkes or Wold processes.

Contact

mp2914@ic.ac.uk