Detecting weak dependence in computer network traffic patterns using higher criticism

Matthew Price-Williams

Imperial College London

Nick Heard

Imperial College London

Patrick Rubin-Delanchy

University of Bristol

Summary.

To perform robust statistical anomaly detection in cyber-security, we must build realistic models of the traffic patterns within a computer network. It is therefore important to understand the dependencies that exist between the large number of routinely interacting communication pathways within such a network. Pairs of interacting nodes in any directed communication network can be modelled as point processes where events in a process indicate information being sent between two nodes. For two processes A and B denoting the interactions between two distinct pairs of computers, called edges, we wish to assess whether events occurring in A trigger events to then occur in B. A test will be introduced to detect such dependence when only a subset of the events in A exhibit a triggering effect on process B; this test will allow us to detect even weakly correlated edges within a computer network graph. Since computer network events occur as a high frequency data stream, we consider the asymptotics of this problem as the number of events goes to infinity, whilst the proportion exhibiting dependence goes to zero, and examine the performance of tests that are provably consistent in this framework. An example of how this method can be used to detect genuine causal dependencies is provided using real world event data from the enterprise computer network of Los Alamos National Laboratory.

Keywords: Computer network, Directed interaction network, Triggering, Higher criticism.

1. Introduction

Anomaly detection methods for cyber-security (see, for example, Lazarevic et al., 2003; Neil et al., 2013; Turcotte et al., 2016) build models from computer network data and then monitor for model deviations in an attempt to identify malicious actors. The computer networks modelled are often large, consisting of tens or hundreds of thousands of nodes, generating millions of network traffic events per day, and therefore postulating a full joint model on the entire network is usually infeasible. As a remedy, many published methods (Heard et al., 2010; Neil et al., 2013; Turcotte et al., 2016) make assumptions of temporal independence for network nodes, to provide analytic tractability. However, this is often an uncomfortable assumption, which might be made more tenable by first combining some of the most correlated units into single entities before proceeding with independence assumptions on a slightly reduced structure. Intuitive examples of correlation between event processes in a computer network include DNS lookups during web browsing, NetFlow traffic between IP addresses belonging to the same domain, or email reciprocity (Perry and Wolfe, 2013).

Considering the arrival times of different types of events in a computer network as separate point processes on $\mathbb{R}^+$, we propose dividing network models into dependent components by detecting triggering effects between the event times of different event processes. In particular, we consider the case where just a subset of the event times of one point process trigger an increase in the rate of arrivals of events in another; for this we exploit a method for combining evidence first proposed by Tukey (1976), called *higher criticism*.

The data used to demonstrate the methods are authentication event host logs obtained from the Los Alamos National Laboratory (LANL) computer network (Turcotte et al., 2017), collected from all LANL computers running the Microsoft Windows operating system. The data consist of over 10 000 user accounts which authenticate on over 15 000 different computers over a period of 90 days. For each event record in these data, a username and event ID is provided; the latter indicates what type of authentication event occurred, such as a network log on, a workstation lock, or an interactive log on. Dependencies between different event types will be uncovered, some of which are

intuitive, providing validation.

The remainder of this article is organised as follows: Section 2 proposes a statistical model and testing framework for detecting dependence between two event processes in a computer network, when only a proportion of events exhibit a triggering effect. Section 3 reviews the higher criticism test statistic of Donoho and Jin (2004) and then Section 4 investigates the boundaries of a detectable triggering effect in waiting times. Section 5 introduces the model used for the background intensity of arrivals in a network traffic sequence and shows how the triggering detection framework from Section 2 can be generalised for any background intensity function. Section 6 demonstrates the methodology on the LANL computer network authentication data.

2. Testing for dependence between point processes

Let A and B be two point processes, with intensity functions $\lambda_A(t)$ and $\lambda_B(t)$ respectively. In the context of cyber-security modelling, each point process may represent the event times of connections between a pair of nodes in a computer network, or the times at which a user performed a specific type of activity on their computer. It is of interest to detect cases where events in A cause a temporary increase in $\lambda_B(t)$. There are two simple models of triggered behaviour that the B process might exhibit:

(a) Single Response Model: Following each A event, $\lambda_B(t)$ is increased until the next B event.
(b) Multiple Response Model: Following each A event, $\lambda_B(t)$ is increased for a fixed time τ (see, for example Rubin-Delanchy and Heard, 2014).

Since very large volumes of computer network data can be collected, it is meaningful to investigate the asymptotic behaviour of different tests. Following Donoho and Jin (2004), we focus on the non-trivial regime where a diminishingly small proportion of the A events influence the intensity of B. To this end we consider a variation of the Single Response Model. For potentially infinite sequences of event times $a_1 < a_2 < \dots$ from process A and $b_1 < b_2 < \dots$ from process B, define the following subsequences of

alternating event times between the two processes: for $i \geq 1$,

$$\bar{a}_i = \begin{cases} a_1, & i = 1, \\ \min\{a_j \mid a_j \geq \bar{b}_{i-1}\}, & i \geq 2, \end{cases} \qquad (1)$$
$$\bar{b}_i = \min\{b_j \mid b_j \geq \bar{a}_i\}.$$

In this restricted view of the event processes, effectively only one process is being observed at any time, in an alternating fashion. For $i \geq 1$, let

$$y_i = \bar{b}_i - \bar{a}_i \qquad (2)$$

be the waiting time for the ith B process event following the ith A process event in this reduced view, illustrated by the thick green lines in Figure 1. It is assumed that the waiting times $\{y_i\}$ are conditionally independent given a model for the background intensity of the B process. The presence of a triggering effect from a subset of the A events on the B process could be detected by finding a corresponding subset of the waiting times $\{y_i\}$ to be significantly smaller than would be expected against the known background rate for the B process.

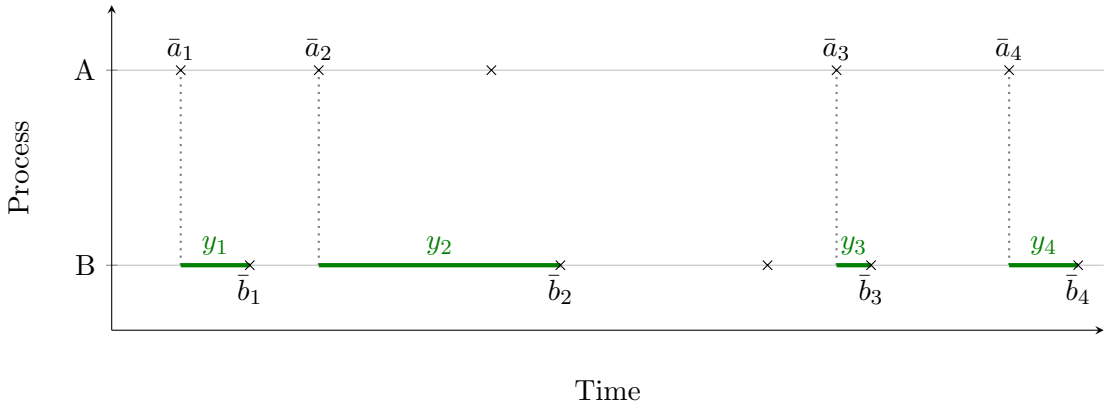


Fig. 1. Example of two event processes A and B. Crosses indicate event times and the solid green lines show the waiting times $y_1, y_2, \ldots$ for the next event in B following an event in A.

Under the null hypothesis of no triggering effect from events in the A process, it will initially be assumed that B is a unit rate homogeneous Poisson process, implying the waiting times (2) are independent realisations from a standard exponential distribution,

denoted Exp(1). This assumption is made only for simplicity of presentation and in Section 5 we show that, using the usual time transformation of counting processes (see, for example, Andersen et al., 2012), this assumption is easily relaxed and the results generalise to more practical null models for $\lambda_B(t)$.

Under the alternative hypothesis of some subset of the A events having a triggering effect, it will be assumed that the corresponding triggered waiting times are from a distribution chosen to yield stochastically shorter waiting times than Exp(1). The simplest choice for the alternative waiting time distribution would be Exp(λ) for $\lambda > 1$. However, in Section 4 we will seek to approximate a transformation of this alternative distribution with a Gaussian distribution, and a more flexible family of models is required. Instead, the triggered waiting times will be assumed to follow a gamma distribution, denoted $\Gamma(\zeta_n, \omega_n)$, with parameters determined according to the sample size n; it will be assumed that the shape parameter $0 < \zeta_n \leq 1$ and the intensity parameter $\omega_n \geq 1$, which guarantees stochastically smaller waiting times than Exp(1).

If ε_n is the long-run proportion of events in $\bar{a}_1, \ldots, \bar{a}_n$ which have a triggering effect, the null and alternative hypotheses are

$$\begin{aligned} H_0 &: y_i \overset{i.i.d}{\sim} \text{Exp}(1), \\ H_1^{(n)} &: y_i \overset{i.i.d}{\sim} (1-\varepsilon_n)\text{Exp}(1) + \varepsilon_n \Gamma(\zeta_n, \omega_n), \end{aligned} \tag{3}$$

A lower tail p-value for the ith waiting time is obtained from the cumulative distribution function,

$$p_i = 1 - e^{-y_i}, \tag{4}$$

such that $p_i \sim \text{Uniform}(0,1)$ under H_0, and under H_1 some proportion of these p-values will be stochastically smaller.

To study asymptotics, following Donoho and Jin (2004) the triggered proportion ε_n in (3) will be imagined to be decreasing in n; against this, the effect size for the triggering implied by the parameter pair (ζ_n, ω_n) will be assumed to be slowly increasing with n.

The aim of this article is to estimate a sequence of distributions of triggered waiting times which lie on the boundary of being asymptotically detectable. As a consequence, any other sequence of distributions (gamma or otherwise) that yield stochasti-

Table 1. An example calculation of the higher criticism test statistic for ten simulated p-values

i	1	2	3	4	5	6	7	8	9	10
$p_{(i)}$	0.005	0.007	0.383	0.438	0.529	0.568	0.792	0.892	0.926	0.964
$\mathrm{HC}_{n,i}$	4.259	**7.320**	-0.540	-0.242	-0.184	0.204	-0.717	-0.937	-0.314	0.611

cally shorter waiting times than these fitted distributions would then also be asymptotically detectable.

3. Higher criticism for Gaussian data

Donoho and Jin (2004) proposed a *higher criticism* (HC) statistic for detecting significant subsets in meta analysis, extending an original idea attributed to Tukey (Tukey, 1976). For a collection of n p-values $p_1, \ldots, p_n$ obtained from n independent significance tests, let $p_{(1)} \leq \ldots \leq p_{(n)}$ be their order statistics. The higher criticism statistic for combining the n tests is

$$\mathrm{HC}_n^* = \max_{1 \leq i \leq n} \mathrm{HC}_{n,i}, \tag{5}$$
$$\mathrm{HC}_{n,i} = \frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})/n}}.$$

For each integer i, $\mathrm{HC}_{n,i}$ takes the proportion of the p-values not exceeding $p_{(i)}$ (by definition, i/n) and subtracts the expected value of this proportion under the null hypothesis ($p_{(i)}$) and then rescales by the standard deviation. Table 1 demonstrates the calculation of (5) for ten randomly simulated p-values, where eight were drawn from the null distribution of U(0,1) and two (0.005 and 0.007) were from an alternative Beta(1,100) distribution, to yield stochastically smaller p-values. The statistics $\{\mathrm{HC}_{n,i}\}$ are maximised at $i = 2$, taking a value of 7.32.

High values of HC_n^* are indicative of a subset of p-values being surprisingly small. Under the null hypothesis, the p-values should all be uniformly distributed on $[0, 1]$, and the null distribution of (5) can therefore be straightforwardly obtained by Monte Carlo simulation. The test static value $\mathrm{HC}_n^* = 7.32$ from the example p-values in Table 1 gives an overall approximate p-value 0.020; from observing ten p-values, it is significant that one fifth of them do not exceed 0.007.

Donoho and Jin (2004) focused on applying (5) to data from a mixture of two unit-variance Gaussians, with null and alternative hypotheses

$$\begin{aligned} H_0 &: x_i \overset{i.i.d}{\sim} \mathrm{N}(0,1), \\ H_1^{(n)} &: x_i \overset{i.i.d}{\sim} (1-\varepsilon_n)\mathrm{N}(0,1) + \varepsilon_n \mathrm{N}(\mu_n, 1). \end{aligned} \tag{6}$$

where $\mu_n > 0$ is increasing in n. For a parameter value $\beta \in (0.5, 1)$ the mixture proportion ε_n was assumed to have a decreasing profile of the form

$$\varepsilon_n = n^{-\beta}, \tag{7}$$

for some $\beta \in (0.5, 1)$.

Let Φ be the standard normal cumulative distribution function. Since $\mu_n > 0$ the likelihood ratio test p-value for the i^{th} test is $p_i = \Phi(1 - x_i)$. Donoho and Jin (2004) obtained boundary conditions for μ_n for which the alternative hypothesis is asymptotically detectable when combining these p-values using (5). Parameterising μ_n as

$$\mu_n = \sqrt{2r \log n}, \tag{8}$$

Donoho and Jin (2004) showed that asymptotically the HC test statistic (5) will distinguish between H_0 and H_1 with probability one provided $r > \rho^*(\beta)$, where

$$\rho^*(\beta) = \min\left\{\beta - \frac{1}{2}, \left(1 - \sqrt{1-\beta}\right)^2\right\}. \tag{9}$$

Consequently, for a given value of β for (7), a sequence of mean effects which are on the boundary of being asymptotically detectable by (5) are given by

$$\mu_n^*(\beta) = \sqrt{2\rho^*(\beta) \log n}.$$

In contrast, Fisher's popular statistic for combining p-values (Fisher, 1925),

$$-2\sum_{i=1}^{n} \log p_i, \tag{10}$$

was shown unable to distinguish between the two hypotheses asymptotically for any $0 < r < 1$ in (8). For the simulated p-values in Table 1, the combined p-value from Fisher's method is 0.124.

The next section will show how the results of Donoho and Jin (2004) for the mixture of Gaussian hypotheses (6) can be translated to the point process dependence application described by the hypotheses in (3).

4. Higher criticism for waiting times

Let $y_1, \dots, y_n$ be the sequence of waiting times drawn from (3). To use the results of Donoho and Jin (2004), we require a transformation of the waiting times $\{y_i\}$ to allow us to link the Exp(1) and N$(0,1)$ distributions, and the $\Gamma(\zeta_n, \omega_n)$ and N$(\mu_n, 1)$ distributions.

A strictly decreasing function for transforming exponentially distributed waiting times $\{y_i\}$ to standard Gaussian variables is given by the composition of the inverse cdf of N$(0,1)$ and the survivor function of Exp(1),

$$x_i = g(y_i) = \Phi^{-1}(e^{-y_i}). \tag{11}$$

If $y_i \sim \text{Exp}(1)$, then $x_i \sim \text{N}(0,1)$. Note that (11) uses the survivor function rather than the distribution function so that large x_i values correspond to shorter waiting times.

Under $H_1^{(n)}$ in (3), together with the transformation (11), $x_i \sim \text{N}(0,1)$ with probability $(1-\varepsilon_n)$, but with probability ε_n, x_i has survivor function

$$\mathbb{P}(x_i > x) = \frac{\gamma\left\{\zeta_n, -\omega_n \log \Phi(x)\right\}}{\Gamma(\zeta_n)},$$

where γ is the lower incomplete gamma function. To translate the asymptotic results of Donoho and Jin (2004) to the waiting time problem by matching distributions, we need to find the parameter pair (ζ_n, ω_n) as a function of μ_n such that $\forall x \in \mathbb{R}$,

$$\frac{\gamma\left\{\zeta_n, -\omega_n \log \Phi(x)\right\}}{\Gamma(\zeta_n)} \approx 1 - \Phi(x - \mu_n). \tag{12}$$

We match the parameters between the two models as follows: Starting with the hypotheses in (3), suppose $Y_0 \sim \Gamma(\zeta_n, \omega_n)$ and $Y_1, Y_2 \overset{i.i.d}{\sim} \text{Exp}(1)$. Setting $Y_1' = Y_1$ and $Y_2' = \min\{Y_1, Y_2\}$, then for $k = 1, 2$,

$$\mathbb{P}(Y_k' > Y_0) = \left(\frac{\omega_n}{\omega_n + k}\right)^{\zeta_n}. \tag{13}$$

For the hypotheses in (6), let $X_0 \sim \mathrm{N}(\mu_n, 1)$, $X_1, X_2 \overset{i.i.d}{\sim} \mathrm{N}(0,1)$. Similarly let $X_1' = X_1$ and $X_2' = \max\{X_1, X_2\}$, and

$$\mathbb{P}(X_k' < X_0) = \int_{-\infty}^{\infty} \{\Phi(x)\}^k \frac{e^{-(x-\mu_n)^2/2}}{\sqrt{2\pi}} dx.$$

To closely match the distributions, we wish to solve

$$\mathbb{P}(Y_k' > Y_0) = \mathbb{P}(X_k' < X_0), \tag{14}$$

for $k = 1, 2$. The case of $k = 1$ can be solved analytically, yielding

$$\mu_n = -\sqrt{2}\Phi^{-1}\left\{\left(\frac{\omega_n}{\omega_n + 1}\right)^{\zeta_n}\right\}. \tag{15}$$

Substituting (15) into (13) and solving (14) for $k = 2$ numerically yields values for the pair (ζ_n, ω_n) which match the relevant tail probabilities of the two distributions. Figure 2 plots the corresponding values of ζ_n and ω_n as functions of $\mu_n \in [0, 5]$.

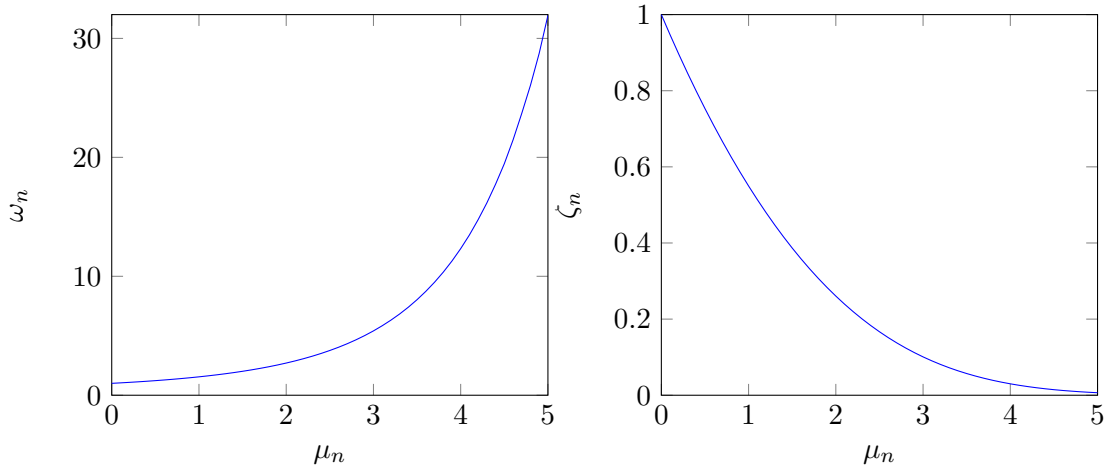


Fig. 2. Estimates of ζ_n and ω_n in terms of μ_n which best match the approximation in (12).

4.1. Illustrative example

For illustration, consider the scenario where $\beta = 0.6$ and $r = 0.2$, implying an effect size which should be should be *just* detectable asymptotically, according to the detection

Table 2. Values of μ_n, ζ_n, ω_n and ε_n for $\beta = 0.6, r = 0.2$

n	μ_n	ζ_n	ω_n	ε_n
10	0.96	0.57	1.52	0.251
100	1.36	0.43	1.86	0.063
1000	1.66	0.34	2.21	0.016
10 000	1.92	0.28	2.57	0.004

boundary (9) since $r > \rho^*(0.6) = 0.19$. Table 2 shows the values of ε_n (7) and μ_n (8) as n increases, and the corresponding values of ζ_n, ω_n from the approximation (12). Figure 3 then compares the corresponding distribution functions from the parameters obtained in Table 2; the fitted gamma distributions are almost indistinguishable from the distributions they are approximating in all cases.

Figure 4 plots the waiting time densities for the null distribution Exp(1), the triggered distribution $\Gamma(\zeta_n, \omega_n)$ and the mixture distribution for the alternative hypothesis in (3). When n is small, a large difference between the densities of the null and alternative hypotheses is required for the alternative hypothesis to be detectable; in particular, the alternative density is much higher for small waiting times. As n increases the density of the alternative hypothesis becomes indistinguishable from the null for larger waiting times and the point at which the curves deviate gets asymptotically closer to zero.

4.2. Simulation study

To demonstrate the effectiveness of the approximation (12) for identifying cases where higher criticism can detect a triggering effect in a subset of waiting times, simulation studies were conducted from the alternative hypothesis in (3). For each $n \in \{10, 100, 1000, 10\,000\}$, n waiting times $y_1, \ldots, y_n$ were drawn from the alternative hypothesis distribution from (3) using the parameter settings from Section 4.1 ($r = 0.2, \beta = 0.6$); the corresponding model parameters for each n were given in Table 2. The lower-tail p-values for each observed waiting $p_i = 1 - \exp(-y_i)$ were obtained and then combined using either higher criticism (5) or Fisher's method (10). This was repeated 10 000 times for each n.

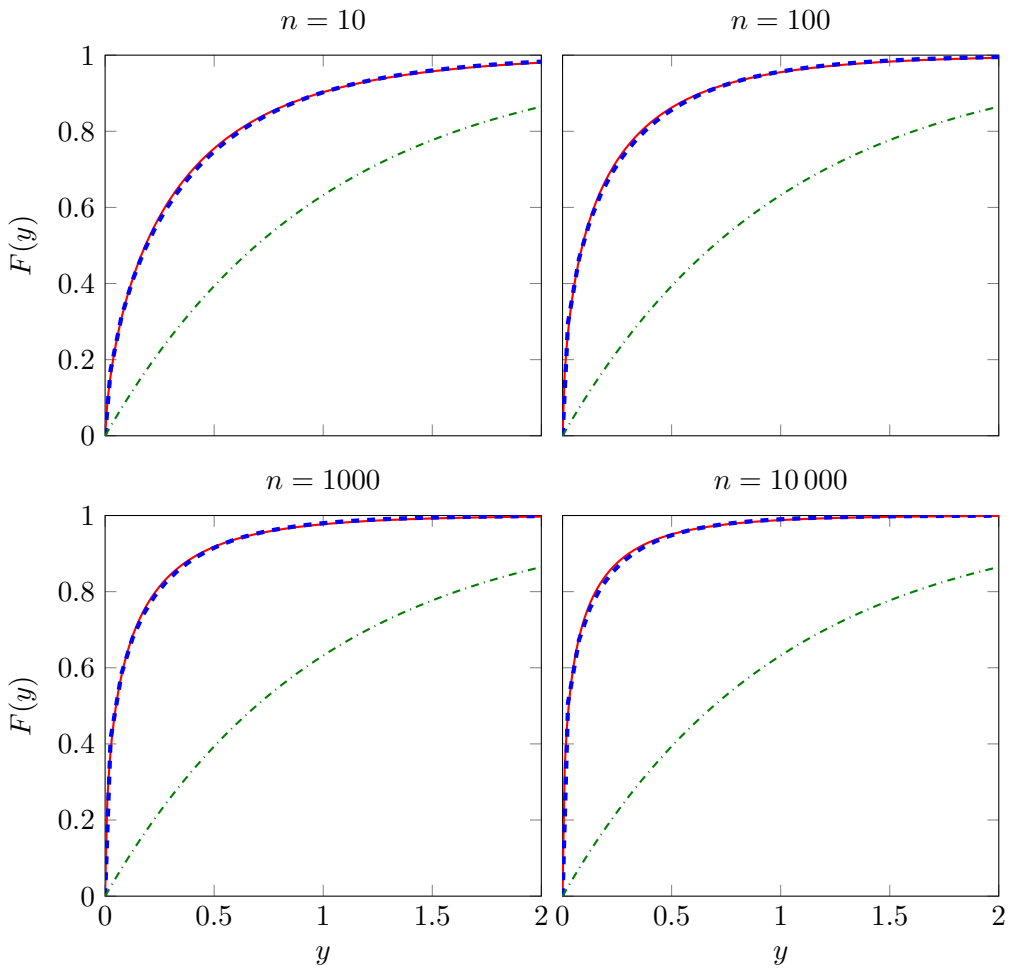


Fig. 3. Distribution function $F(y)$ of waiting times y when distributed $\Gamma(\zeta_n, \omega_n)$ (- - -) compared with the corresponding, transformed $\mathrm{N}(\mu_n, 1)$ (——), and also the null distribution Exp(1) (-·-·-).

Figure 5 presents Q–Q plots obtained from the two methods. For higher criticism, the p-value quantiles converge towards a point-mass at zero as n increases, demonstrating how higher criticism can asymptotically separate the null and alternative hypotheses. In contrast, the p-value quantiles from Fisher's method diverge from those obtained from higher criticism.

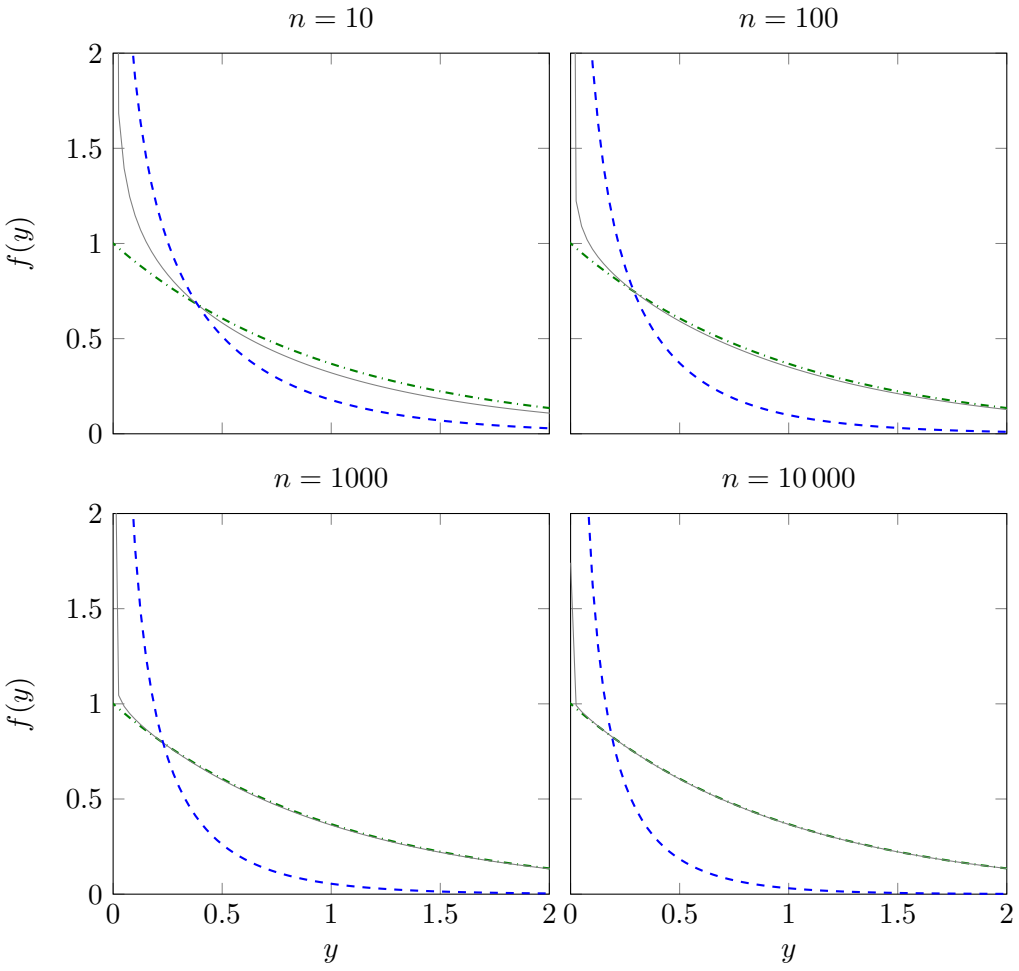


Fig. 4. Density $f(y)$ of waiting times y under the null hypothesis (-·-·-) of (3), the $\Gamma(\zeta_n, \omega_n)$ distribution (- - -) and the alternative hypothesis (——) mixture when $\beta = 0.6, r = 0.2$.

5. Modelling the background intensity

In the waiting time hypotheses (3), the background distribution of events occurring in $B(t)$ is estimated as a unit rate homogeneous Poisson process, implying independent Exp(1) waiting times. Whilst this model is conveniently simple, a more realistic model would capture human user-driven data features such as diurnality/seasonality or self-exciting behaviour, which is commonly present in computer network traffic data where events can be seen to occur in bursts.

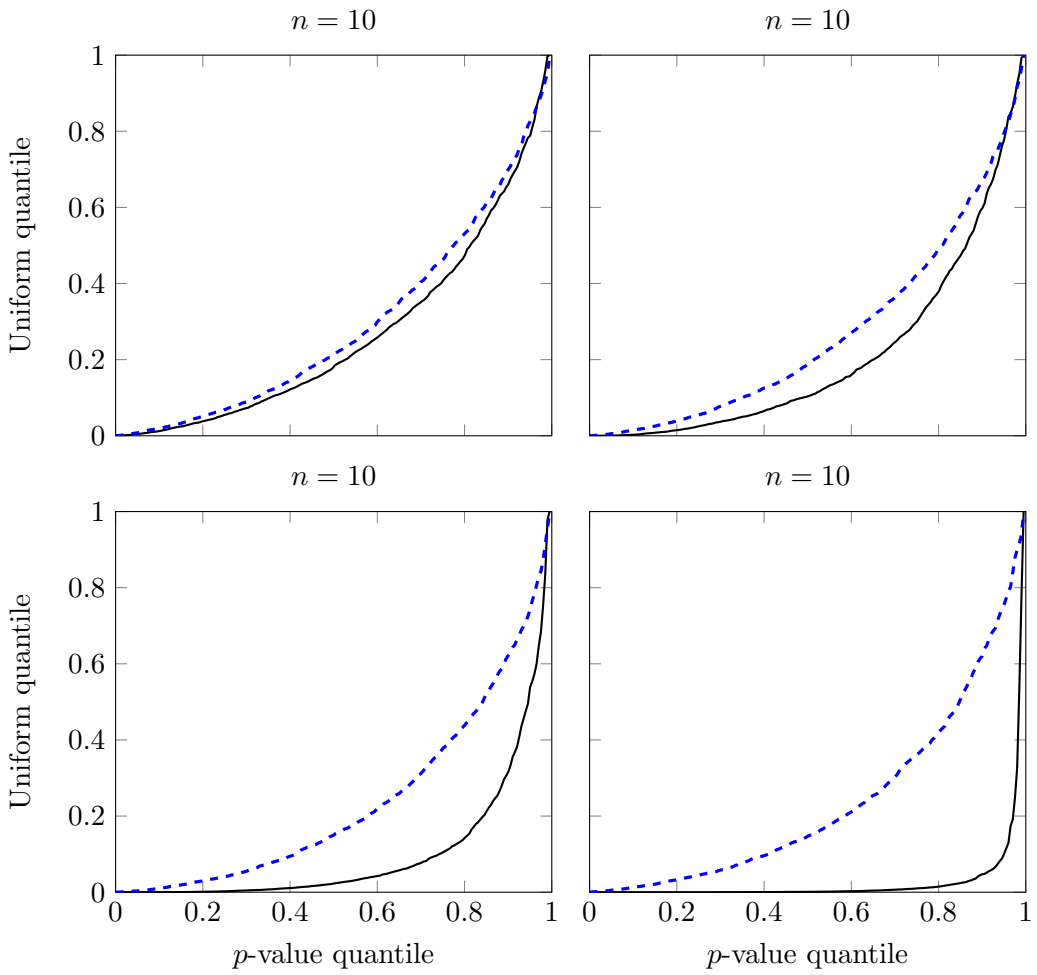


Fig. 5. Q–Q plots of the p-value distributions obtained from higher criticism (——) or Fisher's method (- - -) given n waiting times generated from (3), under H_1 with $\beta = 0.6, r = 0.2$.

Price-Williams and Heard (2017) compared different models for the background intensity of event times in network traffic data, and found best performance was achieved by a Wold process model with a nonparametric excitation function. The intensity function of a Wold process (Wold, 1948) for process B can be written as

$$\lambda_B(t) = \lambda + \omega(t - b^*(t)). \tag{16}$$

where λ is a scalar estimate of the constant background intensity, $b^*(t)$ is the most recent event in B before time t and $\omega(\cdot)$ is a non-negative, non-increasing *excitation* function

which controls the increase and subsequent rate of decay in the intensity following an event in B. Following Price-Williams and Heard (2017), a nonparametric excitation function is specified as a non-increasing step function with an unbounded number of changepoints $\ell \geq 0$. Denoting the changepoints $0 \equiv \tau_0 < \tau_1 < \ldots < \tau_\ell$ and the corresponding step heights as $\lambda_1 > \ldots > \lambda_\ell$, the proposed excitation function is

$$\omega(u) = \sum_{j=1}^{\ell} \lambda_j \mathcal{I}_{[\tau_{j-1},\tau_j)}(u). \tag{17}$$

Details of the estimation procedure for fitting the intensity function (16) with nonparametric excitation (17), applied here using only the event times of the B process to ignore potential triggering effects, are given in Price-Williams and Heard (2017).

This article is concerned with detecting when a subset of the waiting times between events in A and events in B are shorter than would be expected under the background intensity $\lambda_B(t)$. For each event time $\bar{b}_i$ from (1), let

$$y_i = \int_{t=\bar{a}_i}^{\bar{b}_i} \lambda_B(t) dt \tag{18}$$

be the cumulative intensity for the ith waiting time. Note that (18) is simply a generalisation of (2), since the cumulative intensity for standard exponential waiting times is Lebesgue measure.

Under the null hypothesis of no triggering behaviour, the time rescaling theorem (Brown et al., 2002) states that $y_1, \ldots, y_n$ must be exponentially distributed with unit hazard. Under the alternative hypothesis of some subset of the A events having a triggering effect, it will be assumed that the corresponding transformed triggered waiting times are distributed $y_i \sim \Gamma(\zeta_n, \omega_n)$ and hence (3), and the results from Section 4 follow automatically. Note that the background intensity $\lambda_B(t)$ does not necessarily need to be modelled as a nonparametric Wold process, but could instead be modelled by any other appropriate model for the intensity of event times in network traffic data (see, for example Fox et al. (2016); Price-Williams and Heard (2017)).

Table 3. Event IDs used from the LANL authentication data.

Event ID	Description
4624	A user successfully logged onto a computer
4625	A user failed to log onto a computer
4634	A user logged off a computer
4800	The workstation was locked
4801	The workstation was unlocked
4802	The screensaver was invoked
4803	The screensaver was dismissed

Table 4. Log on types used for event IDs $\{4624, 4625, 4634\}$.

Log on type	Description
2	Interactive
7	Unlock
10	Remote Interactive

6. Detecting correlated pairs of event types in authentication data

We now apply triggering detection to the authentication data from the LANL computer network cited in Section 1. The event IDs considered in the analysis are detailed in Table 3, and were chosen as being those most apparently driven by human user interaction, indicated by clear diurnal patterns in the associated event times (see Figure 6). The corresponding log on types for these event IDs are detailed in Table 4.

Two analyses are now presented to demonstrate the methodology. First, higher criticism and Fisher's method are used to detect triggering behaviour between two event IDs across a large selection of users of the LANL computer network. Second, for an example user all correlated pairs of event IDs are detected using higher criticism, partitioning the event types into correlated classes.

6.1. Comparing higher criticism and Fisher's method on authentication data

Intuitively, it seems plausible that a user dismissing their screen saver would often be followed by the same user either logging onto their computer, or failing to log onto their

computer by, for example, inputting an incorrect password. Focusing on the latter case of authentication failure, since this is fairly rare only a small subsection of the screen saver dismissal events should trigger a failed log on event; therefore higher criticism (5) should be better suited to identifying this triggering effect than Fisher's method (10).

For each derived waiting time y_i (18), to test for triggering a lower-tail p-value p_i is calculated for the probability that a random waiting time drawn from the null background intensity is less than or equal to y_i (4). These p-values are combined across the sequence of waiting times using either higher criticism or Fisher's method (10) to obtain an overall level of significance.

To reduce false positives stemming from observing one or two spurious small p-values, Donoho and Jin (2015) proposed a slightly modified version of the higher criticism statistic,

$$\mathrm{HC}_n^+ = \max_{1 \leq i \leq n : p_{(i)} > 1/n} \frac{i/n - p_{(i)}}{\sqrt{p_{(i)}(1 - p_{(i)})/n}}. \tag{19}$$

The statistic (19) disregards very small values $p_{(i)} \leq 1/n$ and therefore only yields large values when a significant subset of the realisations are from an alternative distribution. We choose to use (19) in the present context, since we should expect to observe a reasonable number of triggered events for truly related event types (like the example of a user dismissing a screen saver and then attempting login). To provide a fair comparison, we also adopt an analogously adjusted statistic for Fisher's method,

$$-2 \sum_{1 \leq i \leq n : p_{(i)} > 1/n} \log\{(np_i - 1)/(n - 1)\},$$

to also disregard the very smallest p-values.

The methodology is first applied to the data for an example user from the LANL computer network, User 233172. The left image in Figure 6 plots all event ID 4803 screen saver dismissal and event ID 4625 failed log on events for this user over the 90 days of data collection. Here a red cross represents a screen saver dismissal event and a blue triangle represents a failed log on. Only a reasonably small proportion (46 out of 329) of the screen saver dismissal events are followed by a failed log on event. The right image in Figure 6 plots the 46 p-values calculated for the corresponding waiting

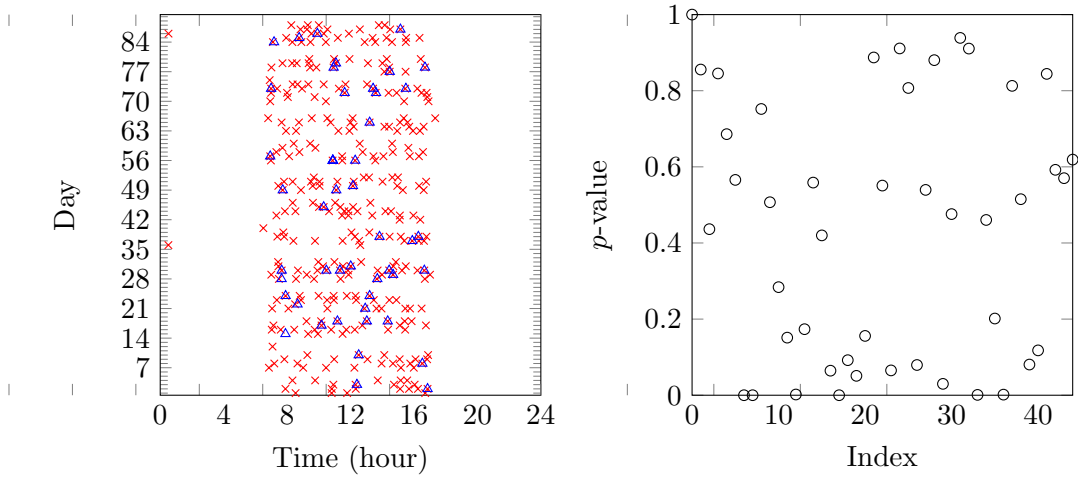


Fig. 6. An example of weak triggering behaviour in authentication data from User 233172 in the LANL network. Left: Times of failed log on events (▵) and screen saver disable events (×). Right: Corresponding p-values for the waiting times between screen saver dismissal and failed log on.

times using the nonparametric Wold process model (16) for the background intensity of the failed log-on events. A small proportion of p-values are very close to zero suggesting triggering for a subset of the events. The combined p-values found using Fisher's method and higher criticism are respectively 0.5032 and 0.0007; higher criticism detects a highly significant triggering effect, whilst Fisher's method does not find any evidence.

This analysis was then repeated for all 1268 users in the LANL network for which at least 20 such waiting times were observed over the 90 days of data. Figure 7 presents a Q–Q plot of the combined p-values calculated using both higher criticism and Fisher's method. The line corresponding to Fisher's method is just below the 45° line implying that Fisher's method does not detect any significant triggering effect between the two event types. In contrast, the higher criticism statistics identifies a triggering effect for the majority of users in the network.

6.2. Detecting correlation between all pairs of event IDs

We now briefly analyse the pairwise triggering relationships between all event ID and log on type pairs for a single user in the LANL network. Taking a graphical perspective,

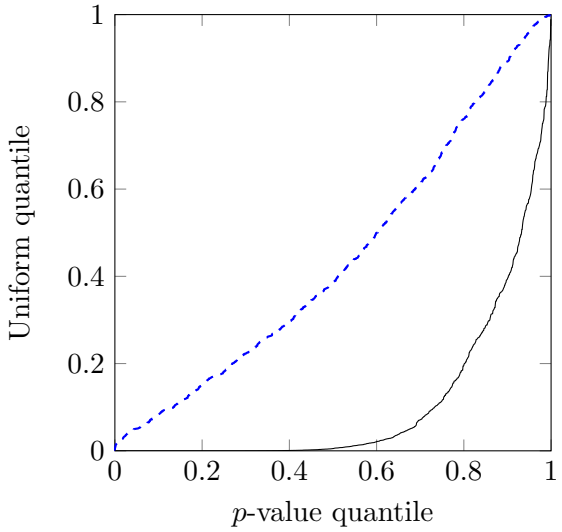


Fig. 7. Q–Q plots of the combined p-values calculated using Fisher's method (- - -) and higher criticism (——) for screen saver dismissal events triggering failed log-on events across 1268 users from the LANL computer network.

each possible pairing of an event ID from Table 3 with a log on type from Table 4 can be viewed as one node; for example event type 4624|7 would correspond to an unlock log on event. We then populate the graph by drawing directed edges between all pairs of nodes where a triggering relationship is detected, using the adapted higher criticism method (19) with nonparametric Wold background intensity (16).

Figure 8 plots the resulting graph for an example user (#236478), after performing triggering analysis with a significance threshold value of 0.001. The methodology partitions the event types into two connected components, with some clear cluster structure; for example, the large, highly-connected cluster contains all four event IDs 480[0-3] concerning lock/unlock or screen saver invocation/dismissal, with no associated log on types; all four of these have a triggering link with one another.

7. Conclusion

This article has investigated statistical methods for detecting correlated traffic patterns in computer networks. A test derived from the higher criticism method of Donoho and

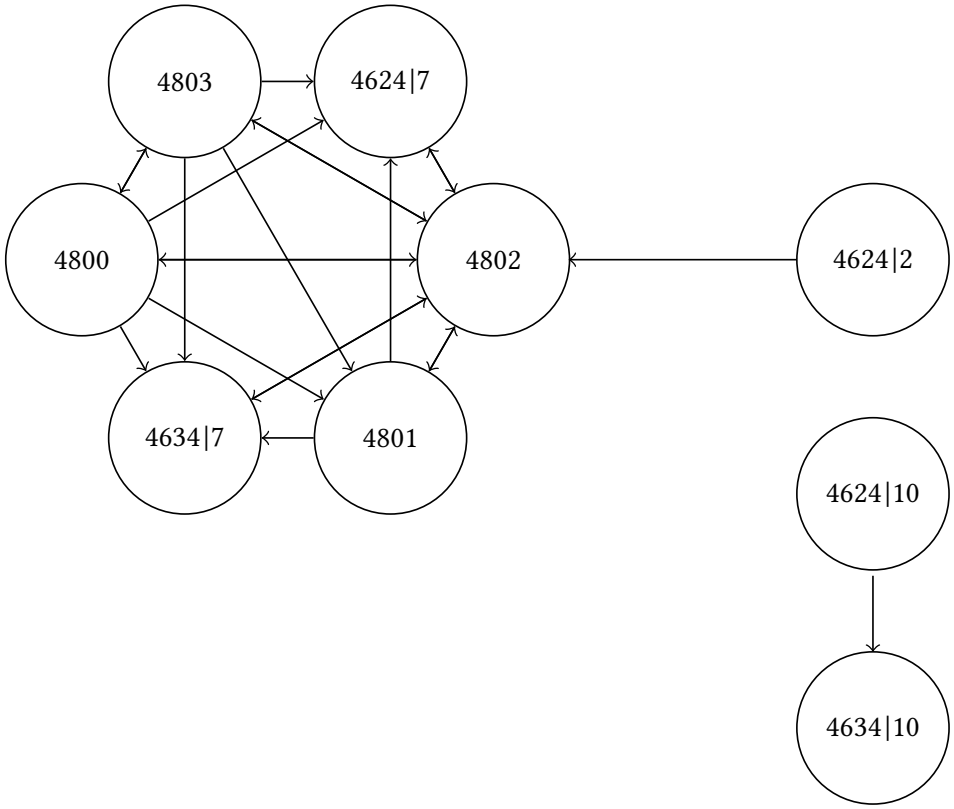


Fig. 8. Diagram of correlated event types for User 236478. A line is drawn between two event types if a dependence is detected between the event times using higher criticism.

Jin (2004) aims to detect triggering effects when possibly only a small proportion of events in one process trigger events in the second process. It was shown that higher criticism was more adapted to detecting this form of triggering than Fisher's method.

It should be noted that there are many other meta analysis methods for combining p-values; see, for example, Heard and Rubin-Delanchy (2018) for a summary of the properties of the most standard methods and when they are most applicable. In particular, Donoho and Jin (2004) notes that Simes' method (Simes, 1986) for combining p-values,

$$\min_i \frac{np_i}{i},$$

shows comparable asymptotic performance to (5) for detecting significant subsets.

An application was presented on authentication data from the LANL computer network where the higher criticism method detected the triggering effect more effectively

than Fisher's method. The test was also used to partition the event types from an example user into smaller correlated sub-networks.

Identifying correlated sub-networks in directed interaction networks is important for reducing false positive rates in anomaly detection. If p-values are to be combined under assumptions of approximate independence, when actually they are highly correlated, then too much significance will be placed on observing several simultaneously small p-values. One spuriously small p-value should not be enough to cross a detection threshold, but this can be reinforced by the correlated p-values which will also necessarily be small, even though they are not really carrying any further information.

A straightforward model for the normal behaviour in a correlated sub-network would combine the event times from all correlated traffic sequences into one counting process before estimating the conditional intensity function of the combined process using one of the methods proposed Price-Williams and Heard (2017). Alternatively, we could construct joint models for each correlated sub-network within the larger computer network. For example multivariate Hawkes or Wold processes Hawkes (1971a,b) could be used to capture the interdependence between event times of event ids in the sub-network.

8. Supplementary Material

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

References

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (2012) *Statistical models based on counting processes*. Springer Science & Business Media.

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

Donoho, D. and Jin, J. (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 962–994.

— (2015) Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.*, **30**, 1–25.

Fisher, R. A. (1925) *Statistical methods for research workers*. Genesis Publishing Pvt Ltd.

Fox, E. W., Short, M. B., Schoenberg, F. P., Coronges, K. D. and Bertozzi, A. L. (2016) Modeling e-mail networks and inferring leadership using self-exciting point processes. *Journal of the American Statistical Association*, **111**, 564–584.

Hawkes, A. G. (1971a) Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, **33**, 438–443.

— (1971b) Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 83–90.

Heard, N. A. and Rubin-Delanchy, P. (2018) Choosing between methods of combining-values. *Biometrika*, **105**, 239–246.

Heard, N. A., Weston, D. J., Platanioti, K. and Hand, D. J. (2010) Bayesian anomaly detection methods for social networks. *The Annals of Applied Statistics*, **4**, 645–662.

Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A. and Srivastava, J. (2003) *A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection*, 25–36.

Neil, J., Hash, C., Brugh, A., Fisk, M. and Storlie, C. B. (2013) Scan statistics for the online detection of locally anomalous subgraphs. *Technometrics*, **55**, 403–414.

Perry, P. O. and Wolfe, P. J. (2013) Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 821–849.

Price-Williams, M. and Heard, N. (2017) Statistical modelling of computer network traffic event times. URL: `http://arxiv.org/abs/1711.10416`.

Rubin-Delanchy, P. and Heard, N. A. (2014) A test for dependence between two point processes on the real line. URL: `http://arxiv.org/abs/1408.3845`.

Simes, R. J. (1986) An improved bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

Tukey, J. W. (1976) T13 n: The higher criticism. course notes. *Statistics*, **411**.

Turcotte, M. J. M., Heard, N. A. and Kent, A. D. (2016) Modelling user behaviour in a network using computer event logs. In *Dynamic Networks in Cybersecurity*, 67–87. Imperial College Press.

Turcotte, M. J. M., Kent, A. D. and Hash, C. (2017) Unified host and network data set.

Wold, H. O. A. (1948) On prediction in stationary time series. *The Annals of Mathematical Statistics*, **19**, 558–567.