Statistical Modelling xxxx; **xx(x)**: 1–27

Renewal model for anomalous traffic in Internet2 links

John Nicholson[1], Piotr Kokoszka[2], Robert Lund[3], Peter Kiessler[1] and Julia Sharp[2]
[1]Clemson University, School of Mathematical and Statistical Sciences, Clemson, SC, USA
[2]Colorado State University, Department of Statistics, Fort Collins, CO, USA
[3]University of California at Santa Cruz, Jack Baskin School of Engineering, Santa Cruz, CA, USA

Abstract: We propose and estimate an alternating renewal model describing the propagation of anomalies in a backbone internet network in the United States. Internet anomalies, either caused by equipment malfunction, news events or malicious attacks, have been a focus of research in network engineering since the advent of the internet over 30 years ago. This article contributes to the understanding of statistical properties of the times between the arrivals of the anomalies, their duration and stochastic structure. Anomalous, or active, time periods are modelled as periods containing clusters or 1s, where 1 indicates a presence of an anomaly. The inactive periods consisting entirely of 0s dominate the 0–1 time series in every link. Since the active periods contain 0s, a separation parameter is introduced and estimated jointly with all other parameters of the model. Our statistical analysis shows that the integer-valued separation parameter and five other non-negative, scalar parameters satisfactorily describe all statistical properties of the observed 0–1 series.

Key words: heavy tails, internet anomalies, on-off process, renewal process, binary data

Submitted 4 May 2020; **revised** 5 October 2020; **accepted** 24 November 2020

1 Introduction

Research presented in this article is motivated by the need to understand statistical properties of the propagation of anomalies observed in the internet traffic. We study data obtained from a nationwide network linking major hubs in the United States. There are many types of anomalies including Distributed Denial of Service attacks or link failures, but this article is not concerned with their detection or classification. Our aim is to construct a statistical model that describes important aspects of the movement of anomalies across the network. We use a suitable database created by other researchers.

There are many potential benefits to understanding statistical properties of the propagation of anomalies over a nationwide network over a long period of time. One

Address for correspondence: Piotr Kokoszka, Colorado State University, Department of Statistics, Fort Collins, CO 80523, USA.
E-mail: Piotr.Kokoszka@colostate.edu

 10.1177/1471082X20983146

is to facilitate the design of network simulators, which are used to validate computer networks before deployment. Another application is to plan the provisioning of hardware and software resources. There has been extensive research on anomaly detection, literally hundreds of articles, so we do not even attempt a review. Chandolla et al. (2009) provide a comprehensive survey of anomaly detection methods in various applications. Tsai et al. (2009) review 55 studies on intrusion detection in internet networks. Bhuyan et al. (2014) comprehensively survey general network anomaly detection methods, systems, and tools, in terms of the underlying computational techniques, while Liao et al. (2013) summarize the network intrusion detection with respect to different network scenarios, from the perspective of system deployments, timeliness requirements, data sources and detection strategies. The anomaly detection techniques and systems in specific network scenarios, for example, wireless sensor networks, Xie et al. (2011), and internet of things, Zarpelao et al. (2017), have been thoroughly reviewed with respect to the distinct characteristics of their network anomalies and detection requirements. Paschalidis and Smaragdakis (2009) consider a spatio-temporal framework for anomaly detection. Kallitsis et al. (2016) describe a hardware–software framework for attack detection that operates on live internet traffic.

In contrast to extensive research on anomaly detection, there is little work on quantitatively describing the propagation of anomalies through a network. A statistical model for anomaly occurrence and duration could enhance network design and performance and help improve network intrusion detection systems. This low level of understanding of the stochastic structure of anomalous traffic must also be contrasted with a profound understanding of the structure of regular traffic over the internet and its subnetworks. The groundbreaking work of Leland et al. (1994) discovered the self-similar nature of such traffic, many elaborations on their work are presented in Park and Willinger (2000). Most models for regular traffic over relatively short time intervals postulate a fractal or multi-fractal structure with normal marginal distributions. More recent references and a comprehensive network-wide predictive model are given in, for example, Vaughan et al. (2013). We show that in contrast to the self-similar, hence strongly dependent, Gaussian time series models used to describe regular traffic, important aspects of anomalous traffic can be well described by independent, but highly non-Gaussian random variables. We build on the work of Bandara et al. (2014) and Kokoszka et al. (2020). Bandara et al. (2014) constructed and described the database we use and presented a preliminary statistical model based on exponential and normal distributions. Using probabilistic and statistical analysis, Kokoszka et al. (2020) showed that light-tailed distributions are not appropriate to describe times between the arrivals of anomalies, and one must use point processes with heavy-tailed interarrival times. In this article, we present a model not just for the times of arrivals of anomalies, but also for their duration and structure.

The remainder of the article is organized as follows. In Section 2, we introduce the database of Internet2 anomalies used and our modelling approach. Section 3 is dedicated to exploration of statistical properties of the data and model formulation. Building on Section 3, we compute model likelihood in Section 4 and estimate model parameters. Section 5 is dedicated to the study of the distribution of the waiting time

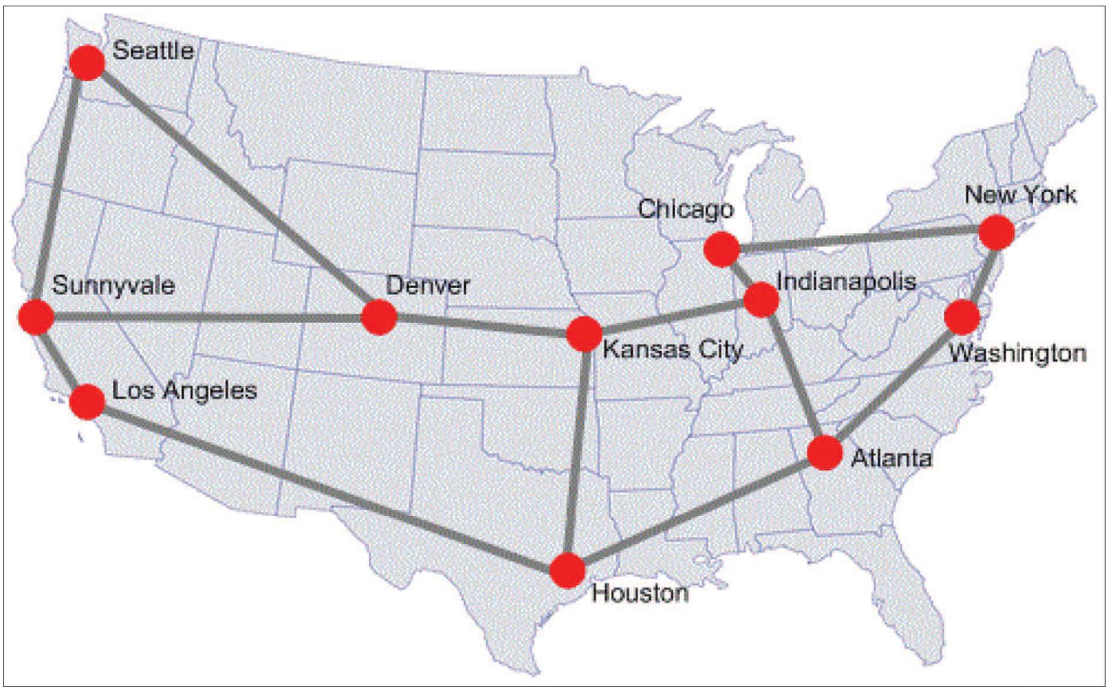

Figure 1 A map showing 14 two-directional links of the Internet2 network
Source: www.internet2.edu

until the arrival of the next anomaly. We conclude with Section 6, where we discuss the limits of our approach and possible future work.

2 Data and modelling approach

We use the database constructed by Bandara et al. (2014) who applied a frequency domain filter to extract time periods of unusually high traffic. Bandara et al. (2014) used traffic measured at the links of the Internet2 network shown in Figure 1 over the period of 50 weeks starting 16 October 2005. Their approach treats periodic and noise components of the measured traffic as usual traffic without anomalies. To extract the anomalies, the 20 largest Fourier components that capture about 80% of the energy and represent the periodic component are removed from the time-series. Then a threshold, between 2 and 3 times the standard deviation of the detrended time-series, is applied. The deviations of the detrended data above or below this threshold are considered anomalous. Generally, if there is an anomaly, the detrended traffic exceeds the threshold by a wide margin, as illustrated in Figure 2. Since a time period of 50 weeks is considered, the final resolution of the temporal records is 5 minutes. For each link, these data can thus be reduced to a string of 0s

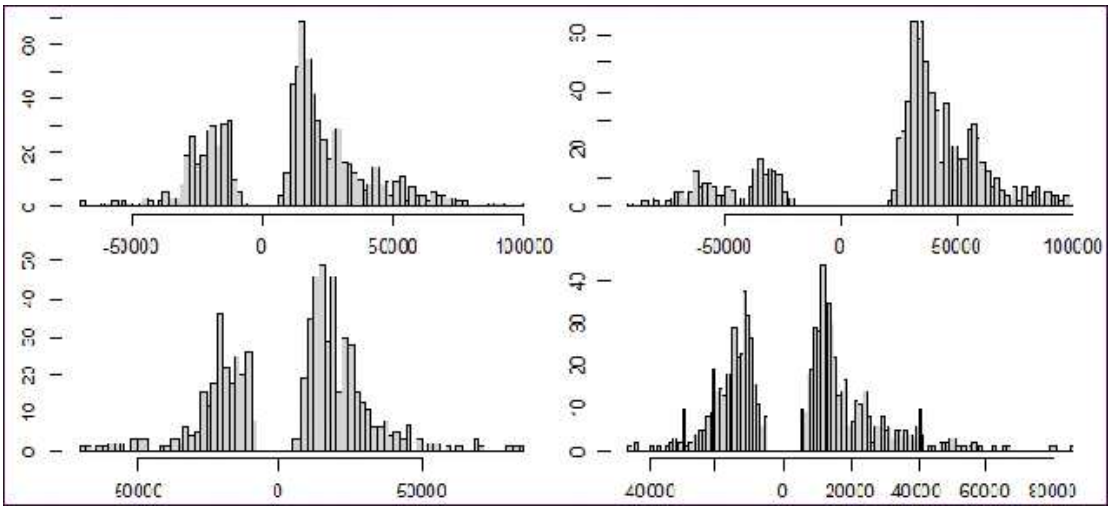

Figure 2 Histograms of the detrended values that are considered anomalous; top-left: incoming Atlanta–Houston, top-right incoming Chicago–Indianapolis, bottom-left: outgoing Denver–St. Louis, bottom-right: the outgoing Houston–Los Angeles

and 1s, where 0 means normal traffic during a five minute-long time interval and 1 means that anomalous traffic occurred during that time interval. There are 14 connections in the graph in Figure 1. Each connection corresponds to two links, for example Seattle→Denver and Denver→Seattle. Measurement devices are installed at the hubs, the nodes of the network. For each link, we thus obtain two slightly different 0–1 strings. For example, for the Seattle→Denver link, we have a 0–1 string coding anomalies leaving Seattle and a different string of anomalies entering Denver. We, thus, have 56 0–1 strings. Unless specified otherwise, the statistical analysis presented below uses the incoming data. The strings are dominated by 0s, a cluster of a few 1s generally occurs after hundreds of 0s. Anomalies are generally separated by days of normal traffic.

Bandara et al. (2014) treat a group of consecutive 1s as a single anomaly, which ends when a 0 occurs. However, examination of the data shows that very often there is a break of just one or two 0s before the next 1. It is reasonable to assume that two strings of 1s separated by a few 0s correspond to a single anomalous event. The issue is then how big a separation should be used to ensure optimal modelling. Using the separation of one recovers the original classification of Bandara et al. (2014). The database does not identify the hundreds of anomalies in the various links by associating them to some exogenously recorded events. We use statistical modelling to determine which separation level leads to a model most likely to explain the observed behaviour of the data.

The data are binary and hence can be modelled as a realization of a random sequence, $\{D_t\}$, of Bernoulli random variables that are neither independent nor identically distributed. Since each string is dominated by 0s, a natural starting point

is to concatenate these long runs of consecutive 0s and record the length of each. Upon doing so, it is clear that the 1s are arriving in clusters and are not individually scattered. The context suggests that it is appropriate to model this time-series by considering any potential time point to fall into one of two categories, those being active periods, where we observe many 1s, and inactive periods, where we find long stretches of 0s in the data. The active periods correspond to anomalies and the inactive periods to regular internet traffic. Thus, we partition the discrete time axis into segments of length X_n, $n = 1, 2, \ldots$. The length of the nth segment is decomposed as $X_n = R_n + A_n$, where R_n is the length of the nth inactive period and A_n is the length of the nth active period. As noted above, a modelling challenge is how to define the active (A) and inactive (R) periods. There are potentially 0s during active periods; if there are many consecutive 0s, it may be suitable to say that the active period has actually ended and the process is in an inactive period. In the definitions that follow, we postulate that an active period has ended at the time after which the process exhibits M consecutive zero. The value of $M \geq 1$ is arbitrary at this point. The statistical analysis that follows will help us determine the optimal range of M for the internet anomalies data. The value of M and other statistical properties of the 0–1 processes will define the statistical model.

Since for each link our data begin with 0, we assume that we start in the middle of an inactive period. For mathematical consistency, we assume that $S_0 = 0$ is the beginning of the first, 0th, (R, A) pair. The event time S_n will be the arrival of the nth (R, A) pair. Formally, we define

$$\begin{aligned} R_1 &= \inf\{k > 0 : D_k = 1\}, \\ A_1 &= \inf\{k > R_1 : D_k = 0, \ldots, D_{k+M} = 0\} - R_1, \\ S_1 &= R_1 + A_1. \end{aligned}$$

We see that R_1 is the time when the first 1 occurs, so R_1 is the length of the first inactive period. We then find the smallest k exceeding R_1 such that it $D_k = 0$, and it is the beginning of a string of M 0s. This is the end of the first active period. After subtracting R_1, we obtain A_1, the length of the first active period. We repeat this process. For $n = 1, 2, \ldots$, we define

$$\begin{aligned} R_{n+1} &= \inf\{k > S_n : D_k = 1\} - S_n, \\ A_{n+1} &= \inf\{k > S_n + R_{n+1} : D_k = 0, \ldots, D_{k+M} = 0\} - (S_n + R_{n+1}), \\ S_{n+1} &= S_n + (R_{n+1} + A_{n+1}). \end{aligned}$$

To illustrate, consider the following (fictitious) data string:

D_k	0	0	1	1	0	1	0	0	0	0	1	1	0	0	0	0	1	1	0	1	0	0
k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

Setting, $M = 2$, we obtain

$$\begin{aligned} &S_0 = 0,\ R_1 = 2,\ A_1 = 4,\\ &S_1 = 6,\ R_2 = 4,\ A_2 = 2,\\ &S_2 = 12,\ R_3 = 4,\ A_3 = 4,\\ &S_3 = 20, \ldots . \end{aligned}$$

Observe that $D_{S_n} = 0$ and

$$S_n = \sum_{k=1}^{n} X_k = \sum_{k=1}^{n} (R_k + A_k), \quad n = 1, 2, \ldots .$$

In the next section, we study distributional and dependence properties of the above model and propose a suitable statistical model.

The modelling approach outlined above can be described as an alternating renewal process. A good introduction to models of this type is given in Section 3.7 of Ross (1996). However, as we will see in Section 3, the commonly used exponential regeneration times are not suitable for the internet anomaly data.

3 Independence and distributional properties of the model

We first analyse the dependence structure of the segments X_n, R_n and A_n. Next we propose models for the distributions of the R_n's and A_ns. Since each active period is allowed to contain zeros, as well as ones, we need to model the distribution of the ones (or zeros) within an active period. Once this is completed, we will have enough information to construct and estimate a likelihood function.

Figure 3 presents plots of the sample autocorrelations of the time-series $\{A_n\}$ and $\{R_n\}$. Examination of analogous plots for other links indicates that it is reasonable to assume that the sequences $\{A_n\}$ and $\{R_n\}$ consist of uncorrelated identically distributed random variables. This conclusion is the same for all values of $1 \leq M \leq 30$. If the sequences $\{A_n\}$ and $\{R_n\}$ each consist of iid observations and are also mutually independent, then the event times $\{S_n\}$ and the companion counting process $\{N(t), t > 0\}$ are known as the alternating renewal process. It is fairly difficult to establish that a given sequence, say $\{Y_n\}$, can be considered a realization of an iid white noise, not just an uncorrelated white noise. An approach established in practice is to compute autocorrelations of transformed observation $f(Y_n)$ for several functions f. If the Y_n are iid, then the $f(Y_n)$ are iid, and hence uncorrelated. We have conducted such an exercise, and determined that the independence assumptions stated above hold to a reasonable approximation. To illustrate, Figure 4 shows the autocorrelations for $\{\log(1 + A_n)\}$ and $\{\log(1 + R_n)\}$. Cross covariances do not show dependence either. We, therefore, proceed with the assumption that $\{A_n\}$ and $\{R_n\}$ are iid sequences independent of each other. In our model, each interrenewal, or interarrival, time $X_n = S_n - S_{n-1}$ is partitioned into 'off' and 'on' periods. The 'off'

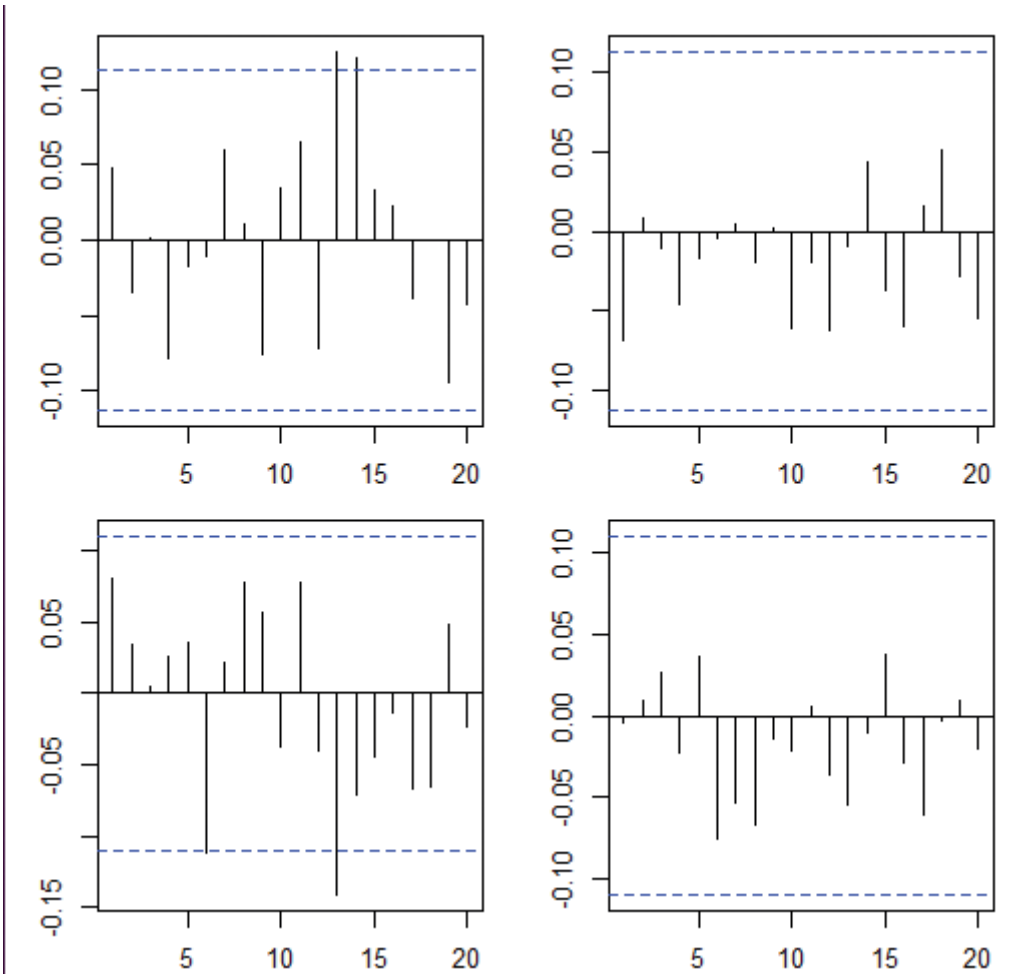

Figure 3 Autocorrelation of lengths of active (left) and inactive (right) periods for the Atlanta→Houston link for $M = 5$. Top panels are for incoming anomalies, bottom for outgoing anomalies. The plots do not suggest any significant autocorrelations. The plots for different values of M look similar

period corresponds to no anomalous traffic and the 'on' period to the presence of an anomaly.

We begin by finding a family of distributions suitable for modelling the length of the inactive periods, the R_n. Since the construction described in Section 2 dictates that $R_i > M$, we fit the distributions to the shifted observations $\tilde{R}_i := R_i - M$. The inactive periods can be very long, and their distribution is definitely heavy-tailed. Our first approach was to fit the discrete Pareto distribution with mass function

$$p(x) = \left[\zeta(\alpha + 1)x^{\alpha+1}\right]^{-1}, \quad x = 1, 2, ..., \alpha > 0,$$

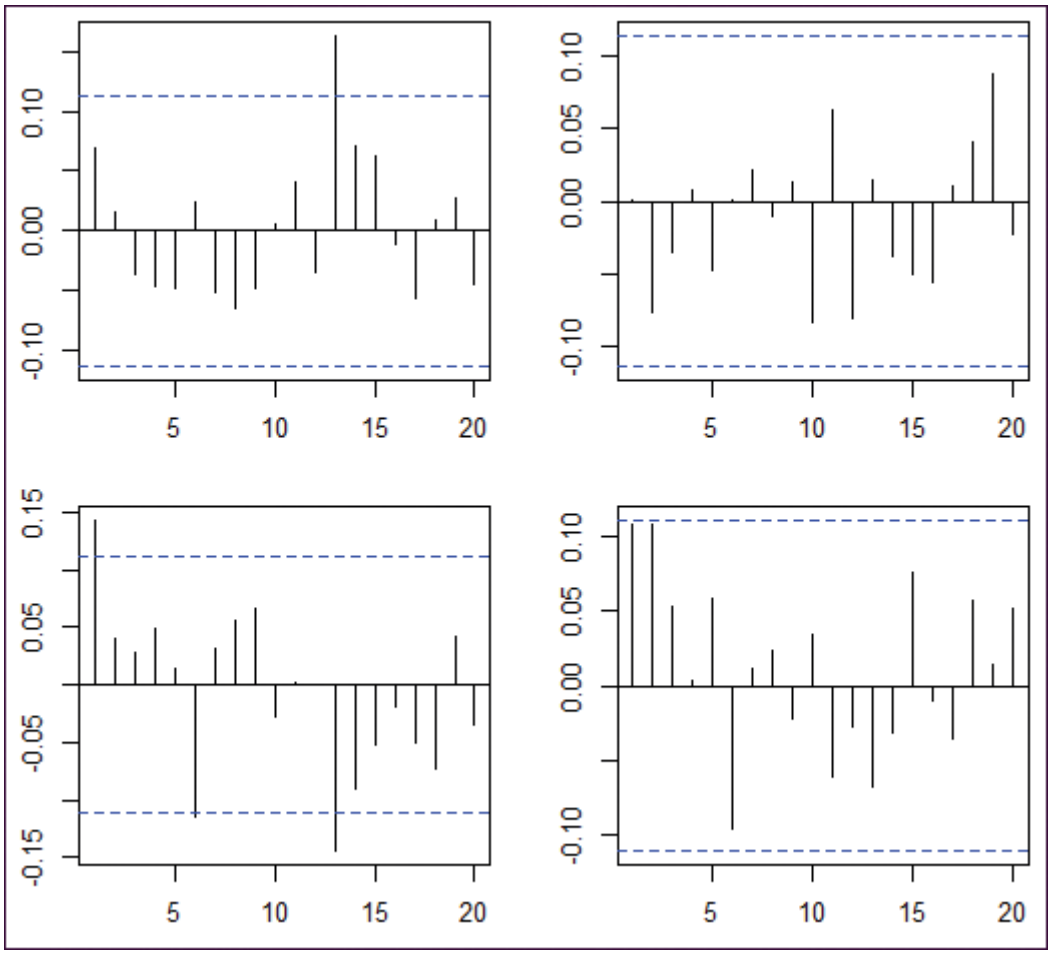

Figure 4 Autocorrelations of natural logarithms of the inputs described in the caption of Figure 3

where ζ is the Riemann Zeta function. However, as seen in Figure 5 this approach resulted in a poor fit. Using a continuous Pareto distribution did not result in any improvement. These distributions lack flexibility in the middle of the data distribution as evidenced by Figure 5. The Pareto Positive Stable (PPS) distribution, see Guillen et al. (2011) and Sarabia and Prieto (2009), is a much more flexible continuous distribution whose distribution function is given by

$$F(x) = 1 - \exp\left\{-\lambda\left[\log(x/\xi)\right]^{\nu}\right\}, \quad x \geq \xi.$$

Note that by taking $\nu = 1$, it reduces to a standard Pareto distribution. There are two ways that we can think of this distribution: the first is by considering it to be the distribution of an exponentiation of a Weibull random variable; so in some sense,

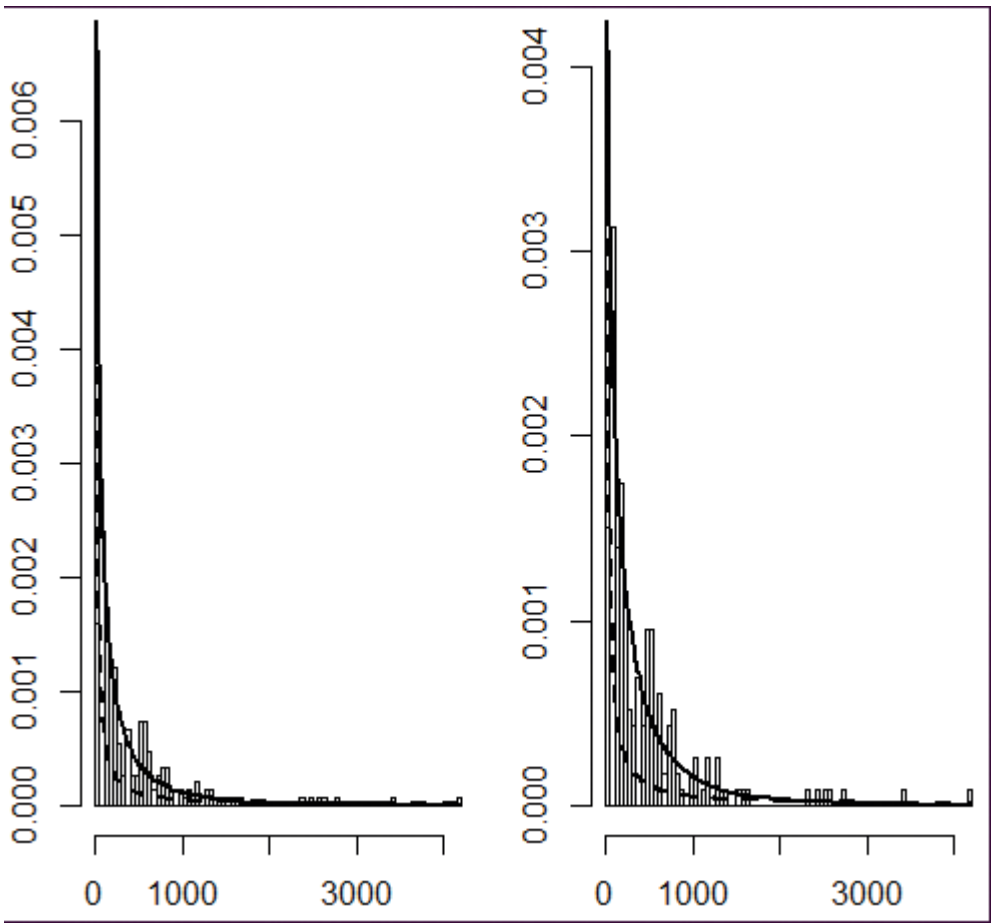

Figure 5 Histograms of the lengths of inactive periods shifted by $M = 5$ (left) and $M = 30$ (right), with the black line being the fitted PPS density, and the dotted line the fitted Pareto density for the Chicago→ NYC link. The PPS density provides a much better fit for the middle portion of the distribution that the Pareto density

we can think of F as being a 'Log-Weibull' distribution. A perhaps more insightful way of thinking about this distribution is to let $X|\alpha \sim \text{Weibull}(\alpha, \xi)$, and let $\alpha \sim G$, where G is a positive stable distribution with Laplace transform $\phi_G(s) = \exp\{-\lambda s^\nu\}$, for $\nu \in (0, 1)$. Then $X \sim \text{PPS}(\lambda, \xi, \nu)$, and so it follows that F can also be seen as a continuous mixture of a Pareto distribution and a positive stable distribution. In Figure 6, visual diagnostics indicate that the PPS distribution fits the data well. We obtained similar plots for other links. The estimated parameters of the PPS distribution depend on the value of M. Table 1 reports estimated parameters for the values of M selected by the likelihood procedure described in Section 4. Application

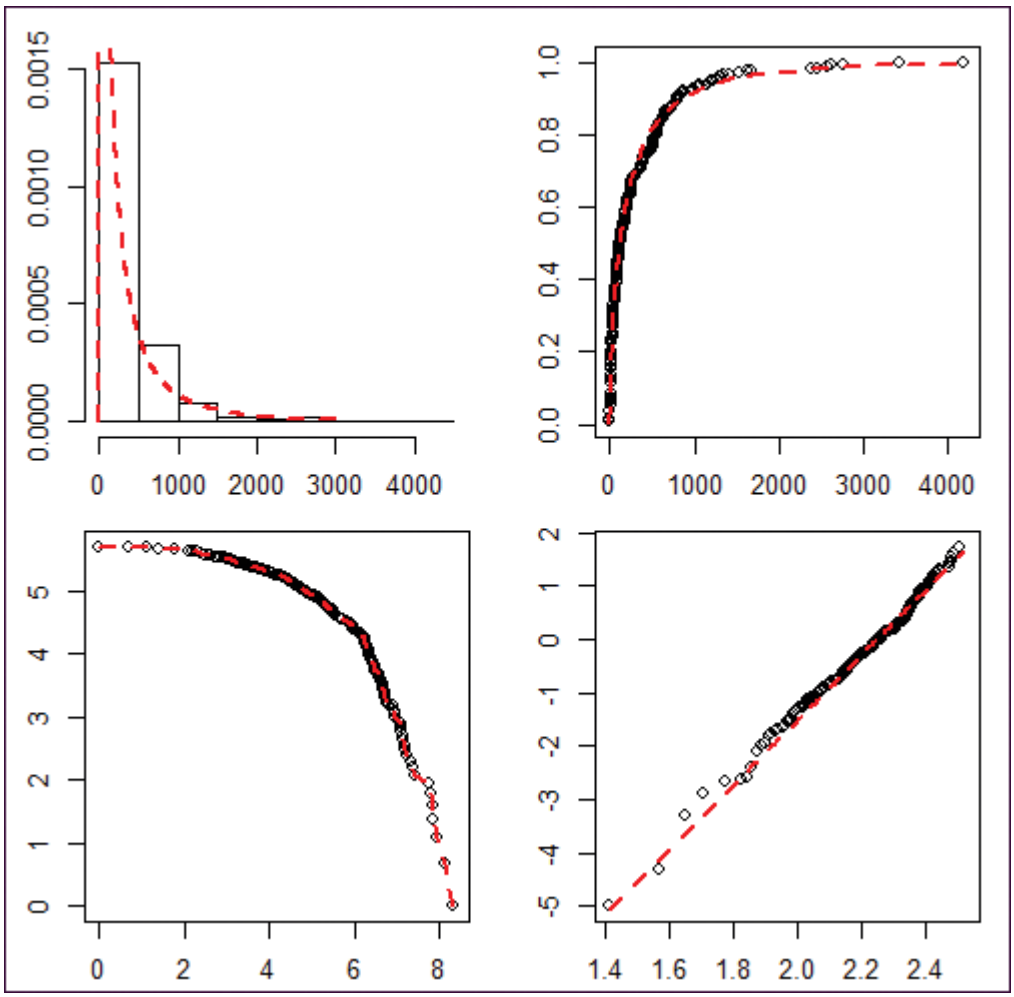

Figure 6 Diagnostic plots visualizing the goodness of fit of the PPS distribution to the data for $R - M$ for the Atlanta→Houston link, and $M = 5$. The first plot gives the fitted PPS density. The second compares the fitted PPS cdf to the empirical cdf. The third is a log-log rank plot which determines tail behaviour (if $R - M$ has a Pareto tail, the line should be straight). The last plot is a double log rank plot which indicates the overall goodness of the data to the PPS distribution

of standard goodness-of-fit tests shows that the the PPS distribution fits well. (These tests strongly reject the exponential distribution.)

We now turn to a model for the active periods. In this case, the discrete Pareto distribution fits the data fairly well visually, as evidenced by Figure 7. However, a potential issue with fitting a Pareto distribution to the active periods is that the maximum observed value can be relatively small for small M. For $M = 30$, the observed values can get fairly large, but for $M = 5$, no value exceeds 25 for the

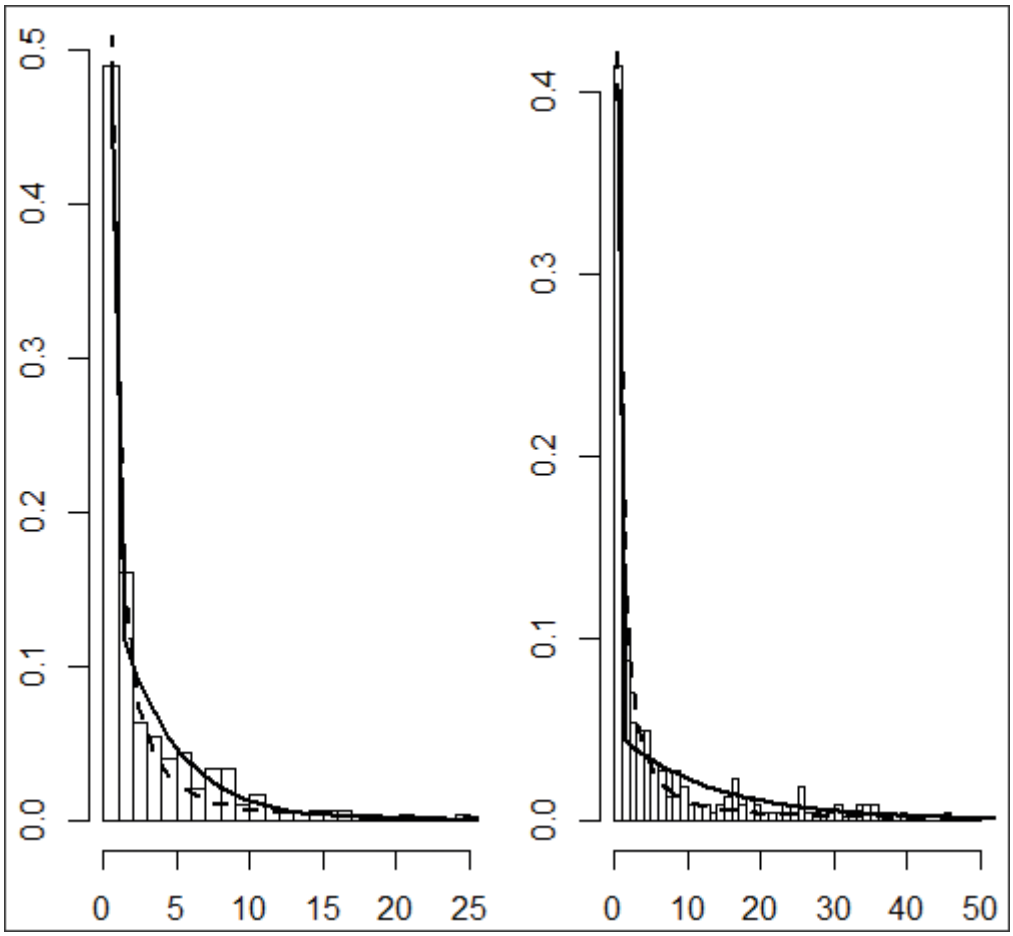

Figure 7 Histograms of the lengths of active periods for $M = 5$ and $M = 30$ for the Chicago→NYC link. The black line is the fitted mixed geometric distribution. The dashed line is the Pareto distribution

data shown in Figure 7, with similar bounds for other links. Therefore, it is not reasonable that a heavy-tailed distribution is a useful model for the active periods. By inspection, we observe that for all links, after a large spike at length $A_n = 1$, the histogram frequencies decrease in a manner that one could argue is geometric. So, we propose modelling the lengths of the active periods with a mixture of a point mass at one and a geometric distribution, that is,

$$p(x; \pi, q) = \pi\delta_1(x) + (1 - \pi)(1 - q)q^{x-1}, \quad x = 1, 2, \ldots, \quad \pi, q \in (0, 1).$$

This distribution, as evidenced by Figure 8, provides a good fit for the data. In addition, the third panel, the log-log rank plot, indicates that the data do not follow a power law, and so the Pareto distribution is inappropriate.

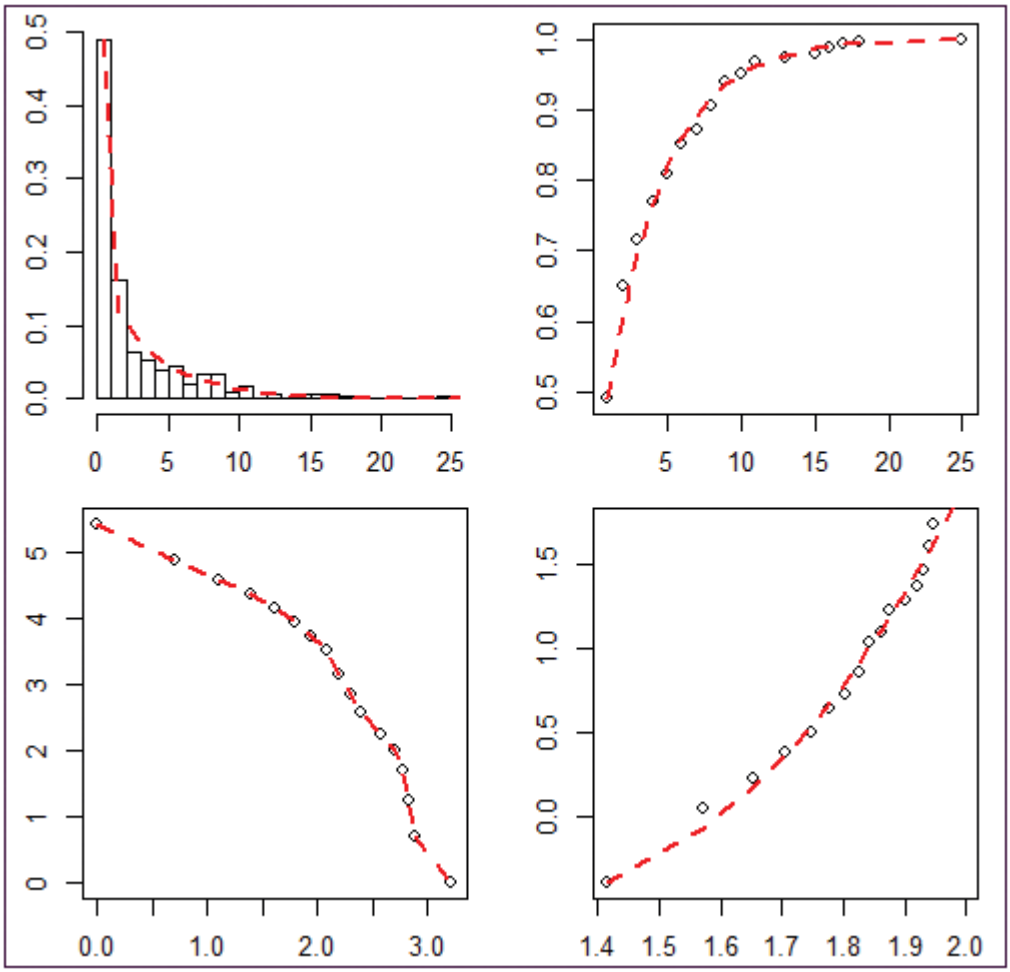

Figure 8 Diagnostic plots visualizing the goodness of fit of the PPS distribution to the data for $R - M$ for the Atlanta→Houston link, and $M = 5$. The first plot shows the fitted mixed geometric density. The second compares the fitted mixed geometric cdf to the empirical cdf. The third is a log-log rank plot which determines tail behaviour (if A has a Pareto tail, the line should be straight). The last plot is a double log rank plot which indicates the overall goodness of the data to the mixed geometric distribution

The remaining piece needed to model the binary string is to model the behaviour of the string during its active periods. Figure 9 shows evidence of a relationship between the length of the active period and the proportion of one's seen during the active period, which is something that should be taken into account. By the construction of the process, we define that the active periods begin after a one has been observed, so it follows that the first element of the string during an active period

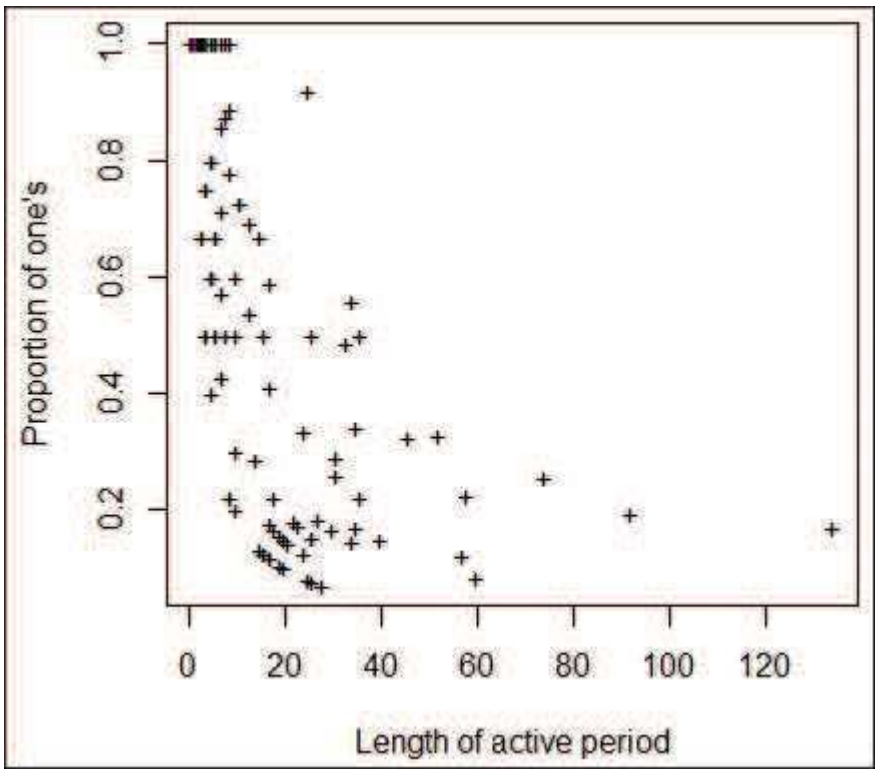

Figure 9 Plot of the proportion of one's during an active period against the length of the active period for the Atlanta→Houston link with $M = 30$

is a one. The high proportion of one's thus follows from construction combined with the fact that there are many anomalies that last less than five minutes. As the length of the active period increases, the proportion of one's decreases, which can also be attributed to how the model is constructed, especially for larger values of M. Thus, we propose a logistic regression model with the predictors being the length of the active period and the current time within the active period. In other words, if we let $\{D_t\}_{t=1}^{A}$ be a binary string representing the values during the activity period, the probability of a one occurring at the tth time during the period is

$$p(t, A) = \frac{\exp\{\beta_0 + \beta_1 t + \beta_2 A\}}{1 + \exp\{\beta_0 + \beta_1 t + \beta_2 A\}}, \quad t = 2, \ldots, A - 1, \quad \beta_0, \beta_1, \beta_2 \in \mathbb{R},$$

where we note that $D_1 = 1$ by our construction. (The time in the logistic regression starts from the beginning of the active period.) We experimented with other models for the probability of 1s as a function of the length of the active periods, but they did not change the likelihoods computed in Section 4 much. So we continue with the logistic regression formulated above.

With all components of the statistical model in place, we turn in the next section to the computation of the likelihood function and parameter estimation.

4 Model likelihood and estimation

To derive the model likelihood, we use the recursions introduced in Section 2 together with the independence properties and distributional models proposed in Section 3. The components of the parameter vector

$$\boldsymbol{\theta} := (\boldsymbol{\theta}_A, \boldsymbol{\theta}_R, \boldsymbol{\theta}_L)$$

are defined by

$$\boldsymbol{\theta}_A = (\pi, q), \quad \boldsymbol{\theta}_R = (\lambda, \xi, \nu), \quad \boldsymbol{\theta}_L = (\beta_0, \beta_1, \beta_2).$$

To understand the principle of constructing the likelihood function, let us consider the toy example of Section 2. Denote by A a random variable with the same distribution as each A_k, and define R analogously. Set

$$p(t, A|\boldsymbol{\theta}_L) = p(t, A|\beta_0, \beta_1, \beta_2) = \frac{\exp\{\beta_0 + t\beta_1 + A\beta_2\}}{1 + \exp\{\beta_0 + t\beta_1 + A\beta_2\}}, \quad 2 \le t \le A - 1.$$

Then,

$$\begin{aligned} &P(D_0 = 0, D_1 = 0, D_2 = 1, D_3 = 1, D_4 = 0, D_5 = 1) \\ &= P(R_1 = 2, A_1 = 4, D_3 = 1, D_4 = 0) \\ &= P(D_2^\star = 1, D_3^\star = 0|A_1 = 4)P(A_1 = 4)P(R_1 = 2) \\ &= p(2, 4|\boldsymbol{\theta}_L)(1 - p(3, 4|\boldsymbol{\theta}_L))P(A = 4)P(R = 2), \end{aligned}$$

where $D_t^\star$ denotes the tth value in an active period. Similarly,

$$\begin{aligned} &P(D_6 = 0, D_7 = 0, D_8 = 0, D_9 = 0, D_{10} = 1, D_{11} = 1) \\ &= P(A = 2)P(R = 4) \end{aligned}$$

because the active period of length 2 has two 1s, which must occur with probability one by construction. For the third interarrival period,

$$\begin{aligned} &P(D_{12} = 0, D_{13} = 0, D_{14} = 0, D_{15} = 0, D_{16} = 1, D_{17} = 1, D_{18} = 0, D_{19} = 1) \\ &= p(2, 4|\boldsymbol{\theta}_L)(1 - p(3, 4|\boldsymbol{\theta}_L))P(A = 4)P(R = 4). \end{aligned}$$

Finally, we have the remainder term, $P(D_{20} = 0, D_{21} = 0) = P(R \ge 2)$. The likelihood function for the toy string introduced in Section 2 thus is

$$L(\boldsymbol{\theta}) = p(2, 4|\boldsymbol{\theta}_L)^2(1 - p(3, 4|\boldsymbol{\theta}_L))^2P(A = 4)^2P(A = 2)P(R = 2)P(R = 4)^2P(R \ge 2).$$

The probabilities $P(A = k)$ and $P(R = k)$ are, respectively, functions of the parameter vectors $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ computed as follows. Since the PPS distribution is continuous and the data are discrete, we discretize the distribution by setting $P(R = 1) = F_{PPS}(1.5)$, and $P(R = n) = F_{PPS}(n + .5) - F_{PPS}(n - .5)$, for each $n \geq 2$, where F_{PPS} is the distribution function of the PPS distribution. The distribution of A is already discrete, so no adjustments need to be made.

We now specify likelihood in the general case. Let n be the length of the string for a specific link and let K be the count of renewals, $K = \max\{k : S_k \leq n\}$, which is a function of the data. Additionally, define $\mathbf{d}$, $\{a_k\}$, $\{r_k\}$ and $\{s_k\}$ to be realizations of $\mathbf{D}$, $\{A_k\}$, $\{R_k\}$ and $\{S_k\}$ respectively. The observed data are $\mathbf{d}$, all other quantities are functions of $\mathbf{d}$ and M. Set also

$$p^\star(t, a|\boldsymbol{\theta}_L) = \begin{cases} p(t, a|\boldsymbol{\theta}_L), & \text{if } D_t^\star = 1, \\ 1 - p(t, a|\boldsymbol{\theta}_L), & \text{if } D_t^\star = 0. \end{cases}$$

For each M, the likelihood function then is

$$L(\boldsymbol{\theta}) = P(R \geq S_K) \prod_{k=1}^{K} \left\{ \prod_{t=2}^{a_k - 1} p^\star(t, a|\boldsymbol{\theta}_L) \right\} P(A = a_k) P(R = r_k).$$

Observe that

$$L(\boldsymbol{\theta}) = L_A(\boldsymbol{\theta}_A|\{a_k\})\ L_R(\boldsymbol{\theta}_R|\{r_k\})\ L_L(\boldsymbol{\theta}_L|\{d_{a_k}\}\{a_k\}),$$

with

$$L_A(\boldsymbol{\theta}_A|\{a_k\}) = \prod_{k=1}^{K} P(A = a_k),$$

$$L_R(\boldsymbol{\theta}_R|\{r_k\}) = P(R \geq S_K) \prod_{k=1}^{K} P(R = r_k),$$

$$L_L(\boldsymbol{\theta}_L|\{d_{a_k}\}\{a_k\}) = \prod_{k=1}^{K} \prod_{t=2}^{a_k - 1} p^\star(t, a|\boldsymbol{\theta}_L).$$

This implies that performing maximum likelihood estimation can be achieved by performing partial likelihood estimation on each individual component, easing the complexity of an optimization routine.

Note that in general the final observation of the data will not be the end of the last interrenewal, so for the likelihood to be complete, we need to consider the likelihood of the observations $(d_{S_K+1}, \cdots, d_n)$. This can be calculated explicitly, but the derivation is quite complicated, and does not change the likelihood function significantly, so it is therefore omitted.

In fact, explicit formulas for the MLEs of the parameters of the distribution can be derived. Consider the distribution

$$\tilde{p}(x;\tilde{\pi},\tilde{q}) = \tilde{\pi}\delta_1(x) + (1-\tilde{\pi})\tilde{q}(1-\tilde{q})^{x-2}\mathbb{1}(x \geq 2),$$

and note that here we recover our original distribution by setting $\pi = \frac{\tilde{\pi}-\tilde{q}}{1-\tilde{q}}$ and $q = \tilde{q}$. Note that it is indeed possible in this case for π to be negative, which does not violate any of the conditions on the distribution if the domain of π is extended, but the interpretation of the distribution is no longer valid. However, given the appearance of the histograms of the active periods this occurrence is unlikely. Though the MLEs for π, q cannot be explicitly calculated, the MLE's for $\tilde{\pi}, \tilde{q}$ can be calculated quite simply. Let $\tilde{L}_A$ be the likelihood function for the transformed parameters. Then,

$$\begin{aligned}
\tilde{L}_A((\tilde{\pi},\tilde{q});\mathbf{x}) &= \prod_{i=1}^{n}\tilde{p}(x_i;\tilde{\pi},\tilde{q}) = \left[\prod_{i:x_i=1}\tilde{p}(x_i;\tilde{\pi},\tilde{q})\right]\times\left[\prod_{i:x_i>1}\tilde{p}(x_i;\tilde{\pi},\tilde{q})\right] \\
&= \left[\prod_{i:x_i=1}\tilde{\pi}\right]\times\left[\prod_{i:x_i>1}(1-\tilde{\pi})\tilde{q}(1-\tilde{q})^{x_i-2}\right] \\
&= \left[\tilde{\pi}^{n_1}(1-\tilde{\pi})^{n-n_1}\right]\times\left[\prod_{i:x_i>1}\tilde{q}(1-\tilde{q})^{y_i-1}\right],
\end{aligned}$$

where we define $n_1 := |\{i : x_i = 1\}|$, and $y_i = x_i - 1, i \in \{i : x_i > 1\}$.

Noting that the first term is the likelihood function of a binary random variable, we know that $\tilde{\pi}_{MLE} = \frac{n_1}{n}$. The second term is the likelihood function of a geometric random variable, and so $\tilde{q}_{MLE} = \frac{1}{\bar{y}} = \frac{n-n_1}{\sum_{i:x_i>1}(x_i-1)} = \frac{n-n_1}{\sum_{i=1}^{n}(x_i-1)}$. By properties of MLE's, $\pi_{MLE} = \frac{\tilde{\pi}_{MLE}-\tilde{q}_{MLE}}{1-\tilde{q}_{MLE}}$ is also a MLE of our original distribution, as is $q_{MLE} = \tilde{q}_{MLE}$.

Lastly, we define

$$\widehat{M} = \text{argmax}_{1\leq M\leq 30}L(\theta;\mathbf{d}),$$

that is, $\widehat{M}$ is the value of M producing the maximum likelihood. Table 1 reports the values of $\widehat{M}$ for each link together with all estimated parameters for this specific value of M. We included only the estimates for the incoming anomalies, the general picture is very similar for the outgoing anomalies. the estimates are very similar. We emphasize, that for the anomalies in an $A \to B$ link, the A(out) and B(in) strings can differ in many positions; 1 is likely to change to 0 because there are mostly zeros in the strings. The parameter estimates are very similar though. Following Bandara et al. (2014), we use the following four-letter abbreviations: Atlanta (atla), Chicago (chin), Denver (dnvr), Houston (hstn), Indianapolis (ipls), Kansas City (kscy), Los Angeles (losa), New York (nycm), Sunnyvale (snva), Seattle (sttl) and Washington DC. (wash).

Table 1 The optimal values of M and parameter estimates for these values for each link for the incoming direction

Link	M	λ	ξ	ν	π	q
atla-hstn	6	3.25E-05	6.61E-02	4.95	0.35	0.22
atla-ipls	3	2.32E-05	2.32E-02	4.80	0.44	0.27
atla-wash	2	3.81E-05	1.85E-02	4.62	0.26	0.29
chin-ipls	3	1.56E-05	2.35E-02	5.10	0.20	0.31
chin-nycm	2	4.16E-05	4.53E-02	4.75	0.29	0.32
dnvr-kscy	2	1.57E-04	2.03E-02	4.02	0.26	0.35
dnvr-snva	3	1.43E-04	3.72E-02	4.28	0.28	0.32
dnvr-sttl	2	1.12E-04	2.42E-02	4.17	0.23	0.31
hstn-atla	3	5.99E-05	6.98E-02	4.69	0.37	0.24
hstn-kscy	4	2.33E-06	1.01E-02	5.64	0.49	0.19
hstn-losa	4	2.57E-05	3.52E-02	4.90	0.32	0.24
ipls-atla	2	7.11E-05	3.13E-02	4.41	0.48	0.30
ipls-chin	3	8.69E-06	2.21E-02	5.35	0.27	0.33
ipls-kscy	2	1.35E-05	1.07E-02	5.01	0.28	0.35
kscy-dnvr	2	3.02E-05	1.94E-02	4.79	0.22	0.39
kscy-hstn	4	3.28E-05	5.99E-02	4.90	0.51	0.25
kscy-ipls	3	1.17E-05	3.92E-02	5.31	0.33	0.31
losa-hstn	3	2.56E-05	2.08E-02	4.81	0.36	0.25
losa-snva	3	7.38E-05	3.34E-02	4.47	0.38	0.28
nycm-chin	3	3.25E-06	1.95E-02	5.75	0.11	0.31
nycm-wash	1	2.21E-05	2.23E-02	4.89	0.25	0.38
snva-dnvr	3	7.63E-05	2.76E-02	4.42	0.36	0.26
snva-losa	3	1.26E-04	3.56E-02	4.32	0.30	0.31
snva-sttl	4	1.28E-05	5.39E-02	5.27	0.77	0.21
sttl-dnvr	5	8.87E-05	9.01E-02	4.45	0.33	0.18
sttl-snva	5	2.60E-05	6.03E-02	5.00	0.63	0.22
wash-atla	3	2.51E-05	4.56E-02	4.96	0.30	0.25
wash-nycm	3	8.76E-06	3.01E-02	5.27	0.26	0.25

We see that the parameter estimates are comparable across all links, so the selected statistical model appears to be appropriate; a misspecified model might work for some links, but not for others. Perhaps the most interesting finding is that the estimated values of ν in the PPS distribution are relatively large. Recall that $\nu = 1$ would correspond to a Pareto distribution, which was used to model tails of $X_n = R_n + A_n$ in Kokoszka et al. (2020) and Kim and Kokoszka (2020). We note that even with the introduction of the separation parameter M, the histograms of X_n and R_n are not very different, especially in the tails, because the active periods A_n are relatively short. For the Pareto distribution, the tail probability is $P(X > x) = cx^{-\alpha}$; for the PPS distribution, it is $P(X > x) = \exp\left\{-\lambda[\log(x/\xi)]^{\nu}\right\}$. For the Pareto distributions, tails become heavier as $\alpha \to 0$; for the PPS distribution if $\lambda \to 0$ and $\nu \to 0$ (ξ is a location parameter). Thus, with the small values of λ in Table 1, the relatively large parameter ν allows us to model the centre of the distribution.

While the other parameters describe distributions, the value of M actually has a physical interpretation in the context of the problem. One can interpret M as the maximum duration of inactivity during an activity period. Since the time difference

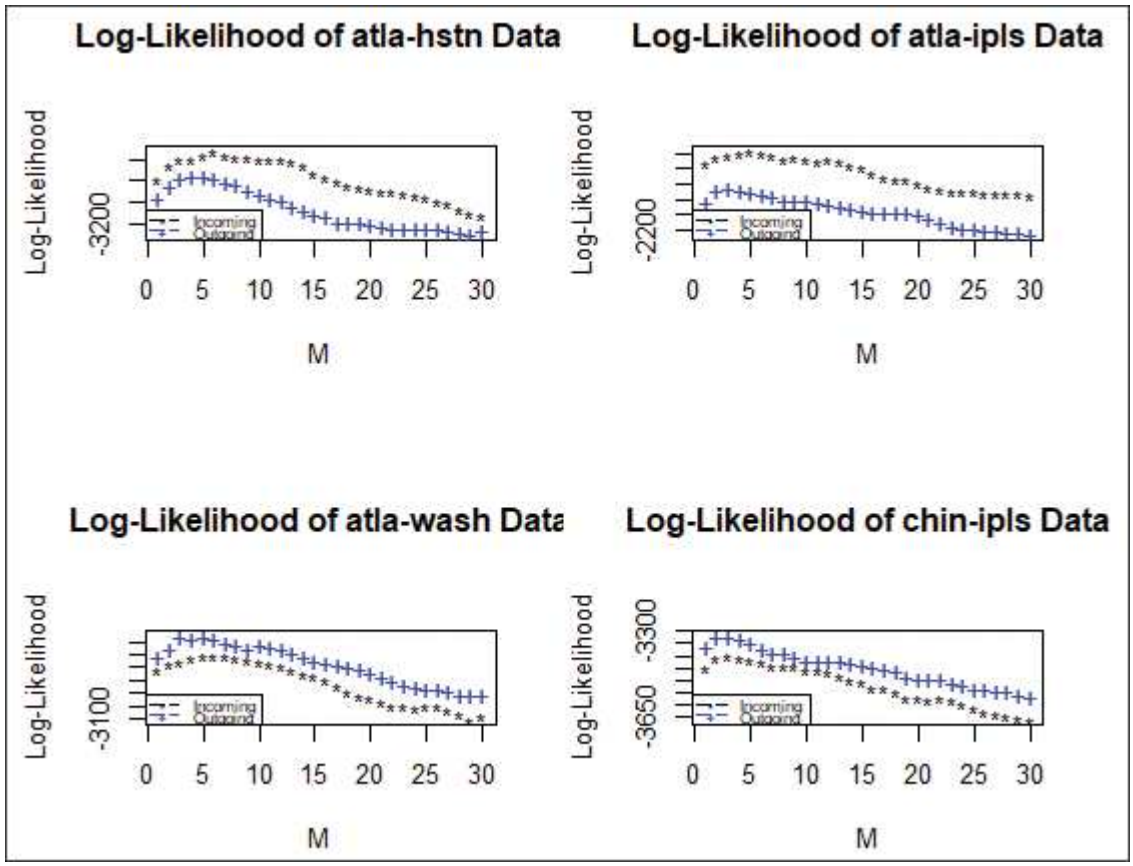

Figure 10 Plots of the log-likelihoods as a function of M for both the Atlanta→Houston link (left) and the Kansas City→Denver link (right). The log-likelihoods for the incoming and outgoing links are shown by '$*$' and '+' respectively. As one can see, the choice of $M = 1$ does not maximize the likelihood for any of the curves, and there appears to be a peak value for the likelihood in M for each

between two points corresponds to five minutes, it would be preferable for M to be reasonably small. Figure 10 shows the log-likelihood functions plotted as a function of M for two links representing cases with a clearly pronounced maximum at M a few units larger than 1, and a weak maximum at $M = 2$. Plots for other links are of one type or the other, or somewhere in between. The maximum is attained at $M = 1$ for only one link. In some cases, the value of M can make a large improvement in the log-likelihood, where as in other cases the improvement is much smaller or not a noticeable improvement at all. Furthermore, after the first several values, the log-likelihoods have a decreasing trend as M increases, which is visible in all of the links. Our modelling approach thus shows that if an inactive period last roughly more than half an hour, one should assume that the anomaly has passed though the link. The explicit values for $\widehat{M}$ are shown for all of the incoming links in Table 1.

However, if we limit our information to that given merely by the interarrivals, and calculate the distribution of $X_n = R_n + A_n$, our model reduces further to the model given in Kokoszka et al. (2020), which provides a distribution as a mixture of a non-negative student-t distribution and a Weibull distribution for X_n explicitly. The model presented in this article does not explicitly state the distribution of X_n, but it is calculated as a convolution of R_n and A_n (since we determined R_n is independent of A_n). Thus, to compare the models in a meaningful way, we will calculate the maximum log-likelihoods for each model across the 28 links.

Given that the model in this circumstance is a basic renewal process, calculating the likelihood function in theory is simple and can be written generally as $L(\theta; \mathbf{x}) = \prod_{k=1}^{K} P(X_1 = x_k)$. From here the likelihood for the model of Kokoszka et al. (2020) can be calculated in a straightforward manner. The likelihood for the model presented in this article employs a Fast Fourier Transform to compute the distribution of $R_n + A_n$. Performing optimization on this convolution did not prove to be particularly stable, but using the parameters estimated for the complete model, the likelihood was an improvement over its counterpart regardless, so solving this problem was unnecessary.

Table 2 gives evidence that our model indeed yields the higher likelihood with the same number of parameters. In addition, our model provides a better framework for taking into account both active and inactive periods separately, whereas in doing any further work along the lines of Kokoszka et al. (2020), we would need to define the distribution of the active periods conditional upon the length of the inter-arrivals, which is nontrivial. Thus, the model where we allow M to vary not only is an improvement in terms of this modelling component but also in terms of the fact that we are performing maximum likelihood estimation on the entire binary sequences rather than a reduction that tells us when the next activity period starts.

5 Distributions of the waiting times

An important application of a statistical model for the distribution of interarrival times in a renewal process is that it can be used to compute waiting times. Denote the current time by t. The anomalies arrive at times $V_k = S_{k-1} + R_k$, $k = 1, 2, \ldots$, so the first anomaly after time t arrives at time $V_{N(t)}$, where $N(\cdot)$ is the counting process defined by $N(t) = \max\{k : V_k \leq t\}$, that is, $N(t)$ is the count of anomaly arrivals up to and including time t. Therefore, the waiting time is

$$W(t) = V_{N(t)+1} - t.$$

Somewhat counter-intuitively, the waiting time is stochastically larger than the interarrival time $V_k - V_{k-1}$. This is because an arbitrary time t has a greater chance of falling into a long interarrival time than a short one, and so there is a higher probability that the time until the next arrival will be long. This effect is particularly well pronounced if the interarrival times are heavy-tailed, that is, long interarrival times occur with a relatively high probability. Waiting times are important for

Table 2 Comparison of the maximum log-likelihoods computed in this article, $\log L_{M,R,A}(\hat{\theta})$, with the maximum log-likelihoods computed using the model of Kokoszka et al. (2020), $\log L_X(\hat{\theta})$

	Link	$\log L_{M,R,A}(\hat{\theta})$	$\log L_X(\hat{\theta})$
1	atla-hstn	−1 984.17	−2 452.70
2	atla-ipls	−1 405.37	−1 709.17
3	atla-wash	−1 917.59	−2 250.87
4	chin-ipls	−2 334.30	−2 746.73
5	chin-nycm	−2 012.67	−2 357.23
6	dnvr-kscy	−1 851.24	−2 120.46
7	dnvr-snva	−2 821.96	−3282.38
8	dnvr-sttl	−1 615.24	−1 924.53
9	hstn-atla	−1 982.83	−2 308.84
10	hstn-kscy	−1 732.67	−2 147.71
11	hstn-losa	−1 855.38	−2 158.22
12	ipls-atla	−1 720.52	−2 020.23
13	ipls-chin	−2 552.88	−3 003.55
14	ipls-kscy	−2 605.01	−3 135.67
15	kscy-dnvr	−2 588.83	−3 109.10
16	kscy-hstn	−1 730.80	−2 037.54
17	kscy-ipls	−2 225.82	−2 597.07
18	losa-hstn	−1 890.74	−2 165.33
19	losa-snva	−2 171.13	−2 522.85
20	nycm-chin	−2 536.60	−3 081.83
21	nycm-wash	−2 011.27	−2 336.96
22	snva-dnvr	−2 085.05	−2 488.28
23	snva-losa	−2 730.21	−3 169.85
24	snva-sttl	−1 718.03	−2 034.74
25	sttl-dnvr	−1 356.06	−1 643.61
26	sttl-snva	−1 858.32	−2 268.92
27	wash-atla	−1 830.40	−2 190.69
28	wash-nycm	−1 518.81	−1 857.49

network design and provisioning of resources. Their distribution was investigated in Kokoszka et al. (2020) using a simpler model containing only anomaly arrival times.

Before comparing distributions derived from our model to those obtained by Kokoszka et al. (2020), we need to explain how the distribution of the waiting time can be computed, see Section 7.4.4 of Pinsky and Karlin (2011), or any other comprehensive textbook on renewal processes. Using the key renewal theorem, one can show that the equilibrium tail probabilities of $W(t)$ are given by

$$\lim_{t\to\infty} P(W(t) > x) = \frac{1}{\tau}\int_x^\infty (1 - G(u))du,$$

where

$$\tau = E[V_k - V_{k-1}], \quad G(u) = P(V_k - V_{k-1} \leq u).$$

Table 3 Estimated 25th, 50th, 75th, 90th and 95th percentiles of the waiting time distribution (first columns) and the interarrival time distribution (in parentheses) for the incoming direction for each link. The table suggests the waiting time distribution is significantly stochastically larger than the interarrival distribution

Link	25th		50th		75th		90th		95th	
atla-hstn	121	(33)	337	(123)	749	(452)	1 543	(804)	2 218	(1 236)
atla-ipls	178	(54)	529	(194)	1 400	(534)	2 728	(1 150)	3 605	(2 064)
atla-wash	115	(55)	298	(199)	707	(442)	1 495	(820)	2 052	(1 286)
chin-ipls	92	(27)	257	(118)	719	(308)	2 182	(564)	3 178	(770)
chin-nycm	107	(50)	289	(165)	760	(350)	1 886	(689)	2 799	(1 026)
dnvr-kscy	136	(20)	355	(130)	832	(419)	1 839	(859)	2 443	(1 228)
dnvr-snva	78	(14)	215	(78)	571	(223)	1 460	(478)	1 981	(674)
dnvr-sttl	156	(47)	427	(196)	1 040	(500)	1 911	(1 093)	2 444	(1 838)
hstn-atla	115	(34)	319	(140)	790	(365)	1 685	(827)	2 313	(1 246)
hstn-kscy	134	(70)	378	(196)	956	(462)	1 957	(953)	2 585	(1 539)
hstn-losa	134	(32)	363	(139)	832	(432)	1 720	(864)	2 372	(1 299)
ipls-atla	139	(33)	372	(166)	814	(511)	1 436	(1 029)	1 969	(1 388)
ipls-chin	83	(35)	220	(120)	497	(295)	1 163	(589)	1 886	(782)
ipls-kscy	79	(40)	215	(116)	529	(286)	1 234	(603)	1 791	(820)
kscy-dnvr	82	(17)	220	(99)	526	(268)	1 377	(561)	2 040	(713)
kscy-hstn	135	(64)	399	(151)	1 045	(448)	2 147	(861)	2 870	(1 480)
kscy-ipls	100	(40)	263	(140)	589	(406)	1 566	(628)	2 242	(860)
losa-hstn	133	(25)	384	(128)	1 009	(352)	2 230	(849)	3 012	(1 259)
losa-snva	111	(24)	301	(121)	861	(339)	2 005	(560)	2 680	(1 078)
nycm-chin	81	(37)	224	(103)	541	(294)	1 626	(556)	2 254	(710)
nycm-wash	107	(42)	286	(160)	695	(364)	1 863	(728)	2 728	(943)
snva-dnvr	118	(16)	322	(103)	832	(352)	1 957	(698)	2 514	(1 054)
snva-losa	83	(20)	230	(90)	606	(254)	1 626	(491)	2 171	(732)
snva-sttl	126	(79)	372	(174)	962	(427)	1 981	(1 049)	2 585	(1 358)
sttl-dnvr	190	(57)	504	(252)	1 129	(661)	1 935	(1 496)	2 397	(2 262)
sttl-snva	123	(61)	354	(157)	952	(386)	1 888	(907)	2 398	(1 478)
wash-atla	124	(56)	344	(163)	791	(445)	1 852	(843)	2 776	(1 268)
wash-nycm	149	(87)	398	(252)	1 005	(560)	2 220	(941)	2 967	(1 547)

We note that the parameters τ and $G(\cdot)$ do not depend on n because the interarrival times have the same distribution. Denoting suitable estimators by $\hat{\tau}$ and $\widehat{G}(\cdot)$, we estimate the cdf of the waiting time by

$$\widehat{F}_W(x) = 1 - \frac{1}{\hat{\tau}} \int_x^\infty (1 - \widehat{G}(u))du.$$

A central issue is to determine which estimators to use. Essentially the only consistent estimator of the cdf $G(\cdot)$ is the empirical cdf $\widehat{G}(\cdot)$ defined by

$$\widehat{G}(x) = \frac{1}{n} \sum_{k=1}^{n} 1\left\{V_k - V_{k-1} \leq x\right\}.$$

Recall that n is the count of anomalies in a specific link. A comprehensive comparison of various estimators of τ presented in Kokoszka et al. (2020) revealed that a very good choice for the internet anomalies data is the estimator which can be derived directly from the empirical cdf $\widehat{G}(\cdot)$ via

$$\hat{\tau} = \int_0^\infty (1 - \widehat{G}(x))dx.$$

Using the above estimators, we computed the quantiles shown in Table 3. Note that since the interarrival lengths depend upon M, the choice of M affects the waiting time distribution. So, to appropriately calculate the distributional estimate, $\hat{M}$ was selected for each link to calculate the interarrival times. Given the distributions for the active and the inactive periods, one may infer the waiting time distribution would behave similarly to the waiting time distribution of the PPS distribution. Note that the hazard function of the PPS distribution converges to zero for $\nu > 1$, which is a significant difference from the hazard function being constant. Thus, one would expect that the waiting times would be significantly stochastically larger than the interarrival times, which is validated by Table 3.

Table 4 compares the sample quantiles of the waiting time distribution for the model presented in this article and the model proposed by Kokoszka et al. (2020). The quantiles are slightly larger for our model. This can be explained by the introduction of the separation parameter of M. Since larger values of M will increase the length of both the active and inactive periods, the quantiles of our model should be larger. Essentially very short inactive periods are eliminated in our approach and treated as parts of active periods. However, the differences are rather small, and may be negligible from point of view of network engineering. This, in a sense, confirms our model, because it essentially agrees with a simpler model in an aspect where a simpler model might be sufficient.

A somewhat unexpected conclusion of the statistical model of Kokoszka et al. (2020) is that the expected waiting time for the arrival of the next anomaly is infinite. While infinite waiting times do occur in various stochastic models, their practical consequences are difficult to quantify and use. We now explain why the waiting time is infinite in the model of Kokoszka et al. (2020) and finite in our model. Denote by W the positive random variable whose distribution is the equilibrium distribution of the waiting time, that is, F_W

$$P(W > x) = \frac{1}{\tau} \int_x^\infty [1 - G(u)]du.$$

Denote by X_G the random variable whose cdf is G, that is, X_G has the same distribution as the interarrival times. Direct verification shows that for $p > 0$,

$$EW^p = \frac{EX_G^{p+1}}{(p+1)\tau}.$$

Table 4 Estimated 25th, 50th, 75th, 90th and 95th percentiles of the waiting time distribution for our model (first columns) and the model of Kokoszka et al. (2020) (in parentheses) for the incoming direction for each link. The quantiles given appear to be slightly larger for our, which due to the inclusion of the separation parameter M into the model

Link	25th		50th		75th		90th		95th	
atla-hstn	121	(118)	337	(334)	749	(743)	1 543	(1 531)	2 218	(2 206)
atla-ipls	178	(175)	529	(526)	1 400	(1 400)	2 728	(2 728)	3 605	(3 605)
atla-wash	115	(112)	298	(294)	707	(701)	1 495	(1483)	2 052	(2 052)
chin-ipls	92	(86)	257	(238)	719	(624)	2 182	(1898)	3 178	(2 965)
chin-nycm	107	(101)	289	(270)	760	(678)	1 886	(1614)	2 799	(2 420)
dnvr-kscy	136	(135)	355	(355)	832	(832)	1 839	(1839)	2 443	(2 443)
dnvr-snva	78	(76)	215	(214)	571	(568)	1 460	(1 460)	1 981	(1 981)
dnvr-sttl	156	(152)	427	(423)	1 040	(1 033)	1 911	(1 898)	2 444	(2 443)
hstn-atla	115	(112)	319	(316)	790	(790)	1 685	(1 697)	2 313	(2 313)
hstn-kscy	134	(131)	378	(375)	956	(944)	1 957	(1 946)	2 585	(2 585)
hstn-losa	134	(131)	363	(360)	832	(826)	1 720	(1 720)	2 372	(2 372)
ipls-atla	139	(138)	372	(372)	814	(814)	1 436	(1 436)	1 969	(1 957)
ipls-chin	83	(81)	220	(217)	497	(497)	1 163	(1 157)	1 886	(1 886)
ipls-kscy	79	(77)	215	(211)	529	(523)	1 234	(1 234)	1 791	(1 780)
kscy-dnvr	82	(82)	220	(220)	526	(526)	1 377	(1 377)	2 040	(2 040)
kscy-hstn	135	(132)	399	(396)	1 045	(1 039)	2 147	(2 135)	2 870	(2 870)
kscy-ipls	100	(98)	263	(261)	589	(586)	1 566	(1 566)	2 242	(2 242)
losa-hstn	133	(131)	384	(381)	1 009	(1 003)	2 230	(2 218)	3 012	(3 012)
losa-snva	111	(109)	301	(298)	861	(849)	2 005	(1 993)	2 680	(2 657)
nycm-chin	81	(76)	224	(209)	541	(485)	1 626	(1 329)	2 254	(2 029)
nycm-wash	107	(101)	286	(267)	695	(624)	1 863	(1 554)	2 728	(2 313)
snva-dnvr	118	(118)	322	(322)	832	(832)	1 957	(1 957)	2 514	(2 514)
snva-losa	83	(80)	230	(227)	606	(595)	1 626	(1 614)	2 171	(2 159)
snva-sttl	126	(124)	372	(369)	962	(956)	1 981	(1 981)	2 585	(2 585)
sttl-dnvr	190	(184)	504	(494)	1129	(1 116)	1 935	(1 946)	2 397	(2 396)
sttl-snva	123	(121)	354	(349)	952	(938)	1 888	(1 874)	2 398	(2 396)
wash-atla	124	(122)	344	(340)	791	(790)	1 852	(1 839)	2 776	(2 751)
wash-nycm	149	(146)	398	(396)	1 005	(1 003)	2 220	(2 230)	2 967	(2 965)

If X_G has Pareto tail, $P(X_G > x) \sim cx^{-\alpha}$ with $1 < \alpha < 2$, as in Kokoszka et al. (2020), then $EX_G^2 = \infty$, implying $EW = \infty$. Our model leads to long waiting times whose expected values are however finite, as we now explain. By the independence of the random variables R and A, $\mathrm{Var}[X_G] = \mathrm{Var}[R] + \mathrm{Var}[A]$. Even without independence, $EX_G^2 \leq 2[ER^2 + EA^2]$, so the expected waiting time is finite if $ER^2 < \infty$ and $EA^2 < \infty$. The random variable A has a geometric tail, so all its moments are finite. A sufficient condition for $ER^2 < \infty$ is $\nu > 1$, see Sarabia and Prieto (2009). For all estimated ν in Table 1, $\nu > 4$, so we can safely conclude that our model implies $EW < \infty$ for all links.

We emphasize that the infinite waiting time following from the work Kokoszka et al. (2020) does not imply that its distribution will necessarily have larger quantiles than the distribution used in this article. To illustrate, if a positive random variable X satisfies $EX = \infty$, then for any $c > 0$, $E[cX] = \infty$, but choosing c sufficiently small, any finite quantile of cX can be made arbitrarily small.

6 Summary and next steps

Our work has focused on developing a statistical model for the propagation of internet anomalies in a US-wide network. The same model applies to all links, the parameters depend on the link. There are several novel elements in our approach that could potentially be useful in similar contexts. First, we showed how to conduct an exploratory analysis of an alternating renewal process to establish distributional and independence properties needed to construct a realistic statistical model. Second, we proposed nonstandard distributional models for the length of the inactive periods. Third, we proposed a regression approach to modelling the proportion of 1s as a response to the length of the active period. Fourth, we showed that the separation between the active periods, M, can be estimated. While probabilistic properties of alternating renewal processes have been studied in-depth, there has been little work on constructing a realistic statistical model with a complete estimation methodology. This is where the novelty and the main contribution of our work lies.

A remaining question is how to describe the interaction between anomalies in various links. Through extensive experiments, we came to the conclusion that this would be very difficult within the framework considered in this article, and generally within a framework of statistical rather than engineering modelling. We now discuss the relevant issues and speculate on possible approaches.

Large hubs, the nodes of the network, play a significant role. Hardware and software placed at each node are designed to deal with anomalies. They never do it perfectly, so some anomalies, generally in a modified form, may travel to connecting links. What happens to an anomaly at a node depends on whether it is detected, and if so, how it is classified. A node can also be a source of an anomaly, for example, if a local network it serves is under attack or fails in some way. No such information is contained in our data. We speculate that due to such factors, there is no apparent connection between anomalous traffic at various links, as illustrated in Figure 11.

As the caption of Figure 11 emphasizes, and as has been noted earlier, for each unidirectional link, we actually have two datasets. This is because no measurements can be made in the link itself, which can be, for example, an optical fibre cable. Measurements are made by servers placed at certain locations in the hubs. Thus, say for an Atlanta→Houston link, we have incoming measurements (in Houston) and outgoing measurements (in Atlanta).

There is a statistical dependence between the incoming-outgoing pairs, as illustrated in Figure 12 The plots suggest that there may be significant correlation between the time series of the incoming-outgoing pairs. The lag structure of this correlation appears to be haphazard, we could not discern any pattern that would apply to all links.

The discussion above highlights the limits of modelling that can be done based on the available 0–1 strings. A more complete model would need to involve the action of the nodes and precise labelling of anomalies. The action at a note could potentially be described by an input–output model with multiple inputs and/or outputs. A model of

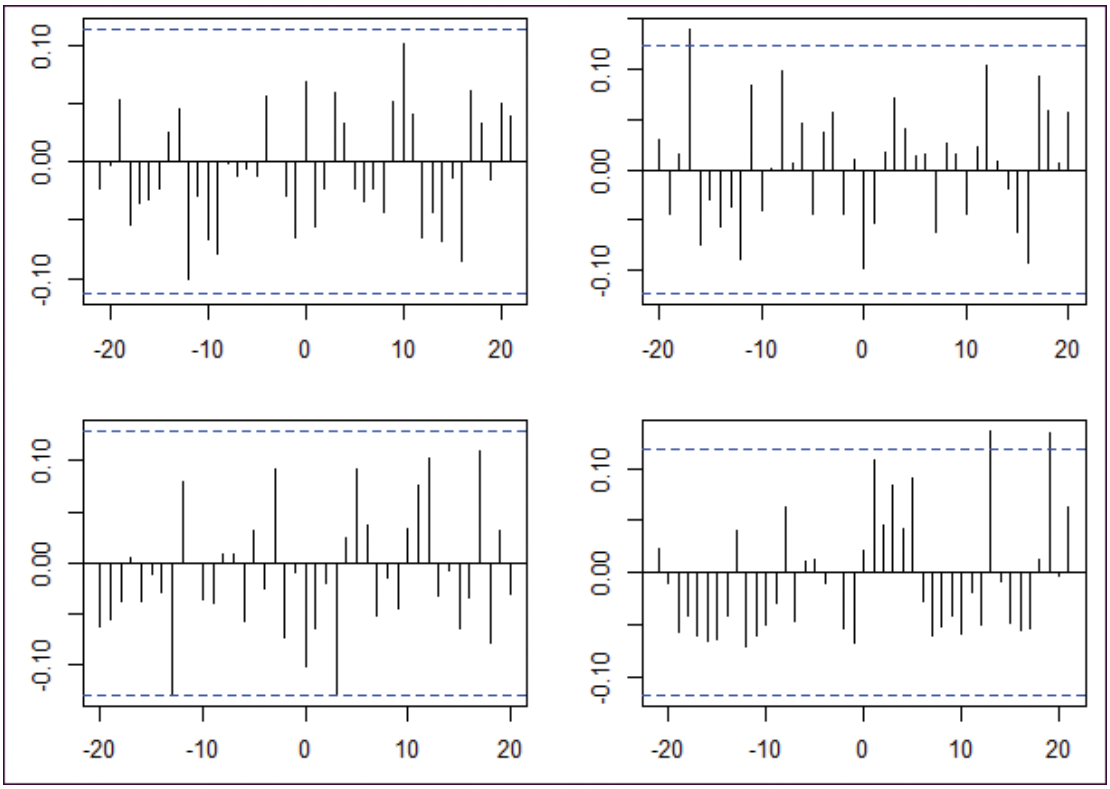

Figure 11 Cross-correlation plots for the interarrival times for the incoming Atlanta→Houston and Chicago→Indianapolis (top-left), incoming Denver→Sunnyvale and Indianapolis→Atlanta (top-right), outgoing Houston→Kansas City and Indianapolis→Atlanta (bottom-left), and outgoing Kansas City→Denver and Sunnyvale→Denver (bottom-right) links. These plots suggest no significant cross-correlation between these different links. From this, one would not expect that incoming and outgoing interarrivals corresponding to distinct links would not possess correlation

this type for brain networks was recently proposed by Sienkiewicz et al. (2017), but it focuses on the node action, and there are no fixed links in the brain. A comprehensive hybrid engineering/statistical model would need to connect the statistical properties of anomalies propagation thorough the links to the action of the nodes. A much more comprehensive and detailed database would need to be constructed before advances in this direction can be made. The model developed in this article could be used to validate any future more comprehensive model, which would need to predict the properties we discovered and quantified.

Finally, it would be of interest to investigate if the model remains valid, or how it would need to be modified, for anomalies extracted from internet traffic over a more recent time period. We hope that our research will motivate network engineers to construct a suitable database on which the model could be tested.

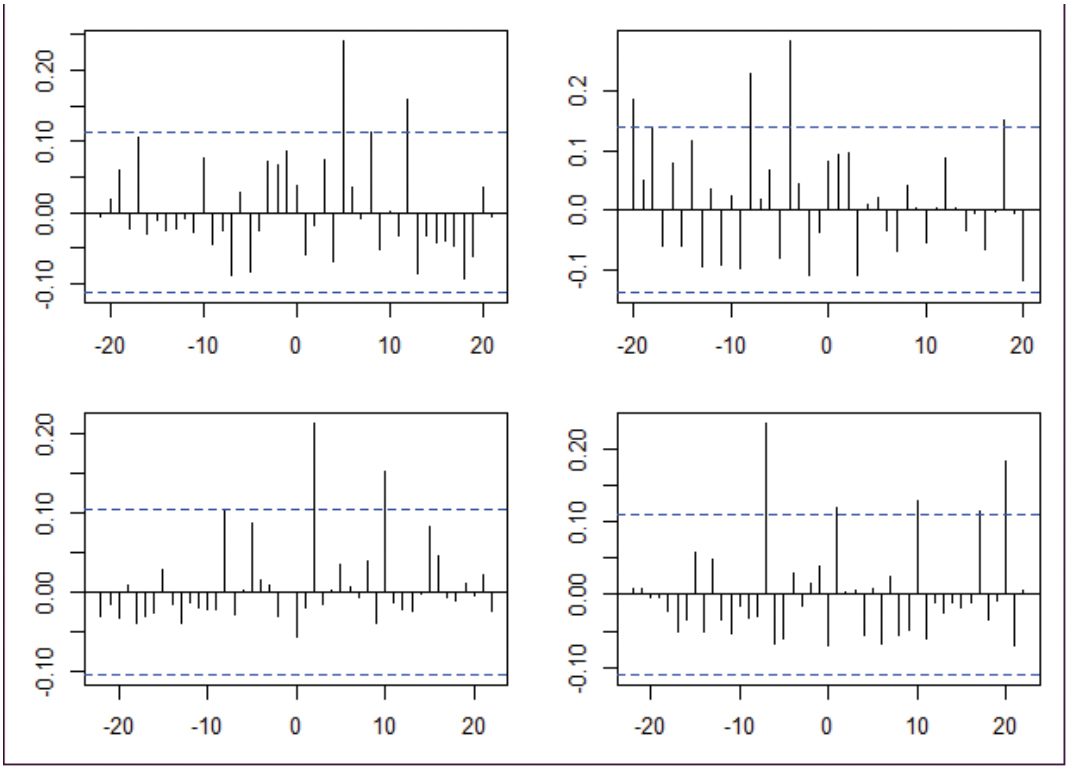

Figure 12 Cross-Correlation plots for the Atlanta→Houston, Atlanta→Indianapolis, Chicago→Indianapolis, and Denver→Sunnyvale incoming-outgoing pairs for $M = 5$

Acknowledgements

We thank Professor Anura P. Jayasumana of the CSU's Department of Electrical and Computer Engineering for sharing the Internet2 anomalies data. We thank the Associate Editor and the referee for reading the article carefully and providing constructive criticism and advice, which helped us to improve the article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This research has been partially supported by NSF grants DMS 1737795 and DMS 1923142.

References

Bandara VW, Pezeshki A and Jayasumana AP (2014) A spatiotemporal model for internet traffic anomalies. *IET Networks*, **3** 41–53.

Bhuyan MH, Bhattacharyya DK and Kalita JK (2014) Network anomaly detection: Methods, systems and tools. *IEEE Communications Surveys & Tutorials*, **16** 303–36.

Chandolla V, Benerjee A and Kumar V (2009) Anomaly detection: A survey. *ACM Computing Surveys*, **41** 15:1–15:58 pages.

Guillen M, Prieto F and Sarabia JM (2011) Modelling losses and locating the tail with the Pareto Positive Stable distribution. *Insurance: Mathematics and Economics*, **49** 454–61.

Kallitsis M, Stoev S, Bhattacharya S and Michailidis G (2016) AMON: An open source architecture for online monitoring, statistical analysis and forensics of multi-gigabit streams. *IEEE Journal on Selected Areas in Communications*, **34** 1834–48.

Kim M and Kokoszka P (2020) Consistency of the Hill estimator for time series observed with measurement errors. *Journal of Time series Analysis*, **41** 421–53.

Kokoszka P, Nguyen H, Wang H and Yang L (2020) Statistical and probabilistic analysis of interarrival and waiting times of Internet2 anomalies. *Statistical Methods & Applications*, **29** 727–44.

Leland WE, Taqqu MS, Willinger W and Wilson DV (1994) On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, **2** 1–15.

Liao H-J, Lin C-HR, Lin Y-C and Tung K-Y (2013) Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, **36** 16–24.

Park K and Willinger W (eds) (2000) *Self-similar Network Traffic and Performance Evaluation*. John Willey & Sons.

Paschalidis IC and Smaragdakis G (2009) Spatio–temporal network anomaly detection by assessing deviations of empirical measures. *IEEE/ACM Trans. Networking*, **17** 685–97.

Pinsky M and Karlin S (2011) *An Introduction to Stochastic Modeling, 4th edition*. Cambridge, MA: Academic Press.

Ross S (1996) *Stochastic Processes*. Hoboken, NJ: Wiley.

Sarabia JM and Prieto F (2009) The Pareto-Positive Stable distribution: A new discriptive model for city size data. *Physica* A, **388** 4179–91.

Sienkiewicz E, Song D, Breidt FJ and Wang H (2017) Sparse functional dynamical models: A big data approach. *Journal of Computational and Graphical Statistics*, **26** 319–29.

Tsai C-F, Hsu Y-F, Lin C-Y and Lin W-Y (2009) Intrusion detection by machine learning: A review. *Expert Systems with Applications*, **39** 11994–12000.

Vaughan J, Stoev S and Michailidis G (2013) Network–wide statistical modeling, prediction and monitoring of computer traffic. *Technometrics*, **55** 79–93.

Xie M, Han S, Tian B and Parvin S (2011) Anomaly detection in wireless sensor networks: A survey. *Journal of Network and Computer Applications*, **34** 1302–25.

Zarpelao BB, Miani RS, Kawakani CT and de Alvarenga SC (2017) A survey of intrusion detection in Internet of Things. *Journal of Network and Computer Applications*, **84** 25–37.