Statistical Methods & Applications (2020) 29:727–744
https://doi.org/10.1007/s10260-019-00500-x

ORIGINAL PAPER

Statistical and probabilistic analysis of interarrival and waiting times of Internet2 anomalies

Piotr Kokoszka[1] · Hieu Nguyen[1] · Haonan Wang[1] · Liuqing Yang[2]

Accepted: 15 November 2019 / Published online: 21 November 2019

Abstract
Motivated by the need to introduce design improvements to the Internet network to make it robust to high traffic volume anomalies, we analyze statistical properties of the time separation between arrivals of consecutive anomalies in the Internet2 network. Using several statistical techniques, we demonstrate that for all unidirectional links in Internet2, these interarrival times have distributions whose tail probabilities decay like a power law. These heavy-tailed distributions have varying tail indexes, which in some cases imply infinite variance. We establish that the interarrival times can be modeled as independent and identically distributed random variables, and propose a model for their distribution. These findings allow us to use the tools of of renewal theory, which in turn allows us to estimate the distribution of the waiting time for the arrival of the next anomaly. We show that the waiting time is stochastically substantially longer than the time between the arrivals, and may in some cases have infinite expected value. All our findings are tabulated and displayed in the form of suitable graphs, including the relevant density estimates.

Keywords Heavy-tailed distributions · Interarrival times · Internet anomalies · Renewal theory

1 Introduction

This paper is motivated by the need to better understand the temporal and spatial structure of anomalous traffic in a nationwide internet network. Characterizing the stochastic structure of the arrivals of Internet traffic anomalies, such as Distributed Denial of Service attacks or link failures, have various practical applications. One is to facilitate the design of network simulators, which are used to validate computer

✉ Piotr Kokoszka
piotr.kokoszka@colostate.edu

[1] Department of Statistics, Colorado State University, Fort Collins, CO 80522, USA

[2] Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80522, USA

networks before deployment. Another application is to predict the arrival of future traffic anomalies and to plan the provisioning of resources. There has been extensive research on anomaly detection. A Google search produces hundreds of papers on anomaly detection in internet or local networks. Chandola 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. Bhuya 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, e.g., 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. Relevant to our research is the work of Paschalidis and Smaragdakis (2009) who consider a spatio-temporal framework for anomaly detection and of Kallitsis et al. (2016) who describe a hardware–software framework for attack detection that operates on live internet traffic.

In contrast to a great deal of attention devoted to anomaly detection, there is practically no work focusing on quantitative description of the propagation of anomalies through the network. In fact, the fundamental quantitative understanding of network anomaly via statistical and probabilistic analysis could significantly enhance the network intrusion performance and shed light on network intrusion detection system design. 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) pointed out to the self-similar nature of such traffic, many elaborations are presented in Park and Willinger (2000). Most models for regular traffic over relatively short time interval postulate a fractal or multi–fractal structure with normal marginal distributions. More recent references and a comprehensive network-wide predictive model are given in e.g. 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 practically independent, but highly non-Gaussian models. We use the framework of heavy-tailed random variables which have been used in many fields of science and economics, see e.g. Adler et al. (1998), Resnick (2007), and Peng and Qi (2017).

Our broad objective in this paper is to contribute to a better understanding of the stochastic structure of anomalous traffic by focusing on a perhaps its most important aspect: the distribution of the arrival times of anomalies. We deploy a set of statistical tools to understand the stochastic structure of the arrival time of Internet2 anomalies over a period of 50 weeks. These tool can be used to study similar problems in other networks. Another question we explore is the distribution of the waiting time for the arrival of the next anomaly. This waiting time is stochastically longer than the time between the arrivals. This is known as the "inspection" or "length-biased sampling" paradox. Intuitively, if t is an arbitrary time instant, it is more likely to fall into a long period between two arrivals than a short period between two arrivals. So the waiting time, starting from time t, until the next arrival can be expected to be longer than the expected time between any two arrivals. We provide a quantitative description of this phenomenon in the context of internet anomaly arrivals.

For the purpose of designing efficient networks, the knowledge of stochastic characteristics of the waiting time may be more important than statistical summaries related to interarrival times. To efficiently design devices or protocols that operate at link exit points, it is useful to know probabilities of an anomaly arriving at any specified future time interval. As explained above, such information is not directly available from physical measurements, as the interarrival times are stochastically shorter than the waiting times for the arrival of the next anomaly. To gain insight about the latter, probabilistic modeling is required.

The statistical methods presented in this paper are different from the usual approaches based on exponentially distributed, memoryless interarrival times, which correspond to Poisson process arrivals. We will demonstrate that the interarrival times are heavy-tailed, i.e. their tail probabilities decay like a power function rather than like an exponential function. This means that the tools based on classic renewal theory must be employed with care, as we attempt to do. We will estimate and examine the characteristics of waiting time distribution and perform hypothesis tests to check if the waiting time distributions are equal among network links and whether waiting time distributions are indeed different from the interarrival time distributions. We will see that they are different, and will explain that this means that the Poisson model cannot be used to describe the anomaly arrival times.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the database of Internet2 anomalies we study. Section 3 is dedicated to exploration of statistical properties of the anomaly interarrival times. Building on the insights obtained in Sect. 3, we estimate in Sect. 4 the distribution of the waiting time for the arrival of the next anomaly. We do it using the tools of renewal theory, whose relevant results are also explained in Sect. 4. The main contributions and finding of the paper are summarized in Sect. 5. The paper contains two appendices which elaborate on the statistical analyses presented in Sects. 3 and 4.

2 Database of anomalies

We use the database of anomalies constructed by Bandara et al. (2014) who applied a simple Fourier transform 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 Fig. 1 over the period of 50 weeks starting October 16, 2005. Their approach treats periodic and noise components of the measured traffic as normal. 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 beyond this threshold are considered anomalous. A group of consecutive anomalous impulses is treated as a single anomaly. The application of their algorithm produces, among other characteristics, a database of anomalies in each link, each described by its start and end times (in 5 min. resolution). In this paper, we

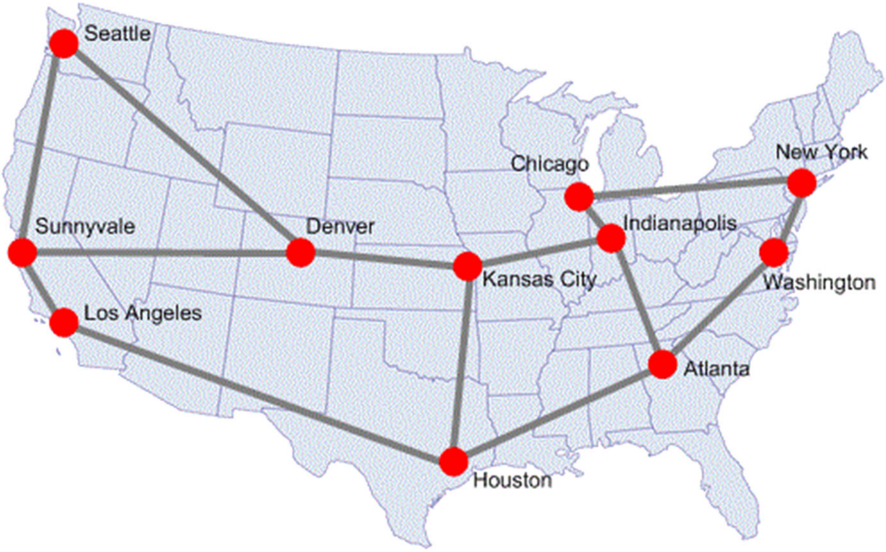

Fig. 1 A map showing 14 two-directional links of the Internet2 network. Source: www.Internet2.edu

will work with interarrival times defined as the time difference between the starts of two consecutive anomalies. We will use 5 min. as the unit lag in the analysis of all time series we consider. In other words, the time separation between generic data points x_t and x_{t+1} is 5 min. An interarrival time is computed as the time difference between the starts of two consecutive anomalies. A typical interarrival time is about 100 lags, i.e. 500 min, but as we shall see in Sect. 3, their lengths vary considerably.

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 D.C. (wash).

3 Statistical properties of anomaly interarrival times

Denote by $S_1, S_2, S_3, \ldots$ the anomaly arrival times in any of the 28 unidirectional links of the Internet2 network shown in Fig. 1. The interarrival times are defined as $X_i = S_i - S_{i-1},\ i \geq 2$. Table 2 shows quantiles of the distribution of these interarrival times, the numbers in parentheses. Figure 2 shows the histograms for selected four links. The visual inspection of the histograms shows that the probability tails decay slower than might be suggested by an exponential distribution; after a peak around zero, the histograms remain fairly flat up to $x = 300$. Simple, rough calculations also show that an exponential decay is not a reasonable assumption, as we now explain. Consider the atla-hstn link in Table 2. The median interarrival time is $m = 53$ and the 95-th percentile is $q = 1056$. Under the exponential model, $e^{-\lambda m} = 0.5$, which gives

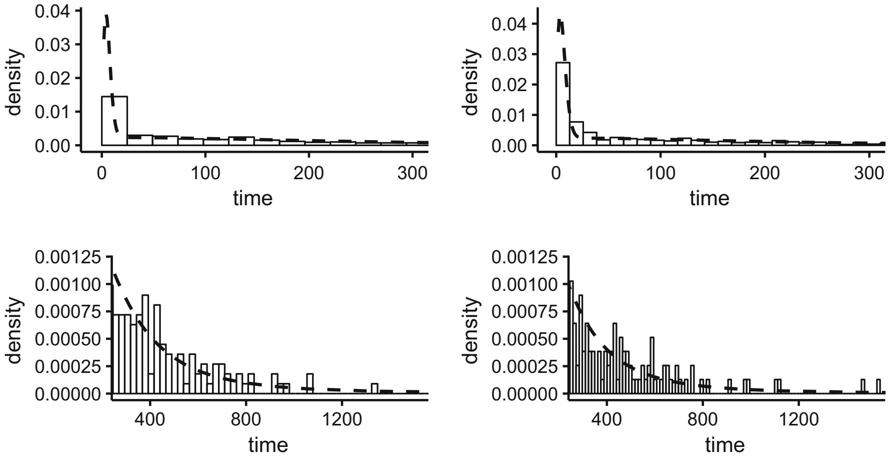

Fig. 2 Histograms of the interarrival times for two ranges with the density of the mixture model considered in Appendix A superimposed: left chin-ipls; right dnvr-snva

$\lambda = 0.0131$. Then q must satisfy $e^{-\lambda q} = 0.05$, which gives $q = 229$, a value about five time smaller than the observed value of 1056. Similar findings hold for other links.

The considerations presented above suggest that the interarrival times might have power-law, or heavy, tails defined by the condition

$$P(X_i > x) \sim cx^{-\alpha}, \quad \text{as } x \to \infty, \tag{3.1}$$

rather than the exponential tails defined by $P(X_i > x) \sim e^{-\lambda x}$. The most common approach to detect heavy tails is to examine a Hill plot, see e.g. Chapter 4 of Resnick (2007). Examples of Hill plots for the interarrival times are shown in Fig. 3. The interpretation of these plots is as follows. If for small values of the “Order Statistics” index the plot levels off, this indicates that relation (3.1) holds. This is what we have observed in the plots for all links. The Hill plot can also be used to estimate the tail index α in (3.1). The value of α is found as the ordinate (alpha) corresponding to the level–off range. Examination of the Hill plots of the interarrival times reveals that it is reasonable to assume that relation (3.1) holds with α generally in the range $1.5 < \alpha < 2$. It is easy to check, using the relation $EY = \int_0^\infty P(Y > y)dy$, that if relation (3.1) holds, then for arbitrarily small $\delta > 0$, $EX_i^{\alpha+\delta} = \infty$ and $EX_i^{\alpha-\delta} < \infty$. Thus, the Hill plots suggest that the expected interarrival times are generally finite, because $\alpha > 1$, but their variance may be infinity because $\alpha < 2$. The latter observation may impact the application of the standard statistical techniques when working with the anomaly interarrival times.

There are several ways of selecting the optimal (in various senses) value of the “Order Statistic” or “Threshold”, which should be used to estimate α. A method that is often used was introduced by Hall (1990). It employs a bootstrap procedure to choose a threshold that minimizes the asymptotic mean square error. This procedure is implemented by the function `hall` in the `R` package `tea`. An example of a Hill

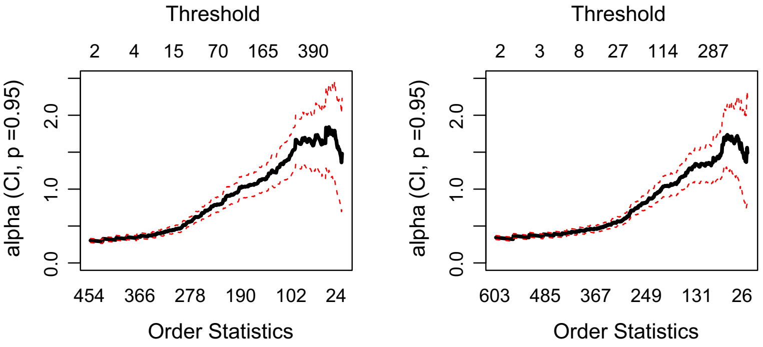

Fig. 3 Hill plots for the interarrival times: left chin-ipls, right dnvr-snva

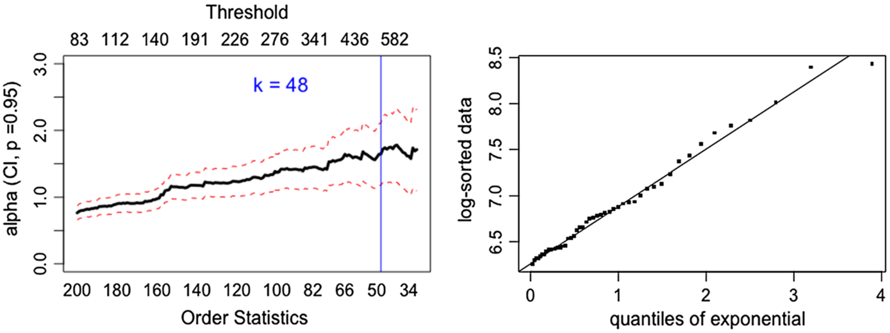

Fig. 4 Hill plot (left) and QQ plot (right) for link 5, chin-nycm

plot with the “Order Statistic” k selected by this method is shown in the left panel of Fig. 4. Table 1 shows the estimated αs. In addition to the Hill estimator, we included three other estimators recommended by Resnick (1997). The values of Hill estimator and the QQ estimator are similar and indicate that the values of alpha are broadly close to $\alpha = 1.5$. The smooth Hill estimator produces smaller values and the moment estimator larger values. All these estimates however support the conjecture that the expected value exists, but the variance may not exist.

We have also explored the validity of model (3.1) by another set of plots, which complement the Hill plots. An example of such a plot is shown in the right panel of Fig. 4, for the link from Chicago to New York (link 5). This is the QQ plot of the log transformed X_i matched against exponential quantiles beyond the exceedance threshold corresponding to the optimal k. We should get approximately a line whose slope is $1/\alpha$ if the data had a Pareto tail with index α, see Section 4.6.4 of Resnick (2007). The QQ plot looks linear with the fit of a straight line whose slope is $1/1.53$, which tells us that it is reasonable to assume a Pareto tail (relation (3.1)) with index 1.53. The QQ plots for other links, with the estimates taken from Table 1, also look linear.

Having established that model (3.1) with $1 < \alpha < 2.5$ is reasonable, we turn to the examination of dependence between the X_i. We will demonstrate the it is reasonable

Table 1 Estimated tail indexes α of the anomaly interarrival times estimated using the data-driven threshold of Hall (1990)

	Link	Hill	Moment	QQ	SmooHill
1	atla-hstn	1.69	1.85	1.50	1.28
2	atla-ipls	1.50	4.19	1.57	1.56
3	atla-wash	1.62	2.00	1.46	1.01
4	chin-ipls	1.62	1.94	1.45	0.98
5	chin-nycm	1.53	1.91	1.50	1.31
6	dnvr-kscy	1.59	1.94	1.51	1.27
7	dnvr-snva	1.68	2.07	1.50	1.23
8	dnvr-sttl	1.56	2.30	1.52	1.48
9	hstn-atla	1.47	2.08	1.51	1.34
10	hstn-kscy	1.44	2.16	1.51	1.35
11	hstn-losa	1.79	1.85	1.51	1.40
12	ipls-atla	2.22	2.19	1.52	1.54
13	ipls-chin	2.11	2.03	1.49	1.21
14	ipls-kscy	1.93	2.30	1.52	1.48
15	kscy-dnvr	2.07	1.85	1.51	1.40
16	kscy-hstn	1.48	2.08	1.51	1.34
17	kscy-ipls	1.91	1.86	1.48	1.11
18	losa-hstn	1.35	2.08	1.51	1.34
19	losa-snva	1.27	1.81	1.34	0.66
20	nycm-chin	1.97	1.85	1.50	1.28
21	nycm-wash	1.97	1.91	1.51	1.42
22	snva-dnvr	1.46	1.90	1.44	0.97
23	snva-losa	1.65	1.86	1.48	1.11
24	snva-sttl	1.43	2.12	1.52	1.36
25	sttl-dnvr	1.83	2.29	1.52	1.55
26	sttl-snva	1.43	2.30	1.52	1.48
27	wash-atla	1.95	2.22	1.52	1.47
28	wash-nycm	1.69	2.16	1.51	1.35

to assume that, for every link, the X_i are independent and identically distributed (iid) random variables. Such assertions are often checked by the examination of the sample autocorrelation function, the ACF. We will follow this approach, but first we emphasize two important points, see e.g. Chapter 1 of Shumway and Stoffer (2017):

1. If the autocorrelations are not significant, this indicates that the observations are *uncorrelated* rather than independent.
2. The significance bands in the ACF plots assume that the fourth moments exists, $EX_i^4 < \infty$.

As explained above, for the anomaly interarrival times, it is even questionable that the second moment exists. To overcome these two difficulties, we will examine ACF plots for transformed X_i, i.e. for $g(X_i)$. The rationale is as follows. If the X_i are independent, then for any g, the $g(X_i)$ are independent as well, and hence uncorrelated. So if we do

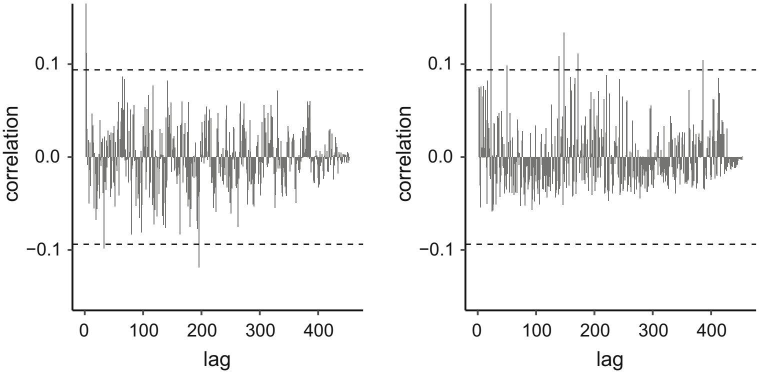

Fig. 5 ACF plots (zero lag excluded) for selected transformations of the interarrival times in the chin-ipls link: left $\ln(X_i)$, right X_i^{-3}

not see significant autocorrelations for several functions g, it is reasonable to assume that the X_i are independent; otherwise significant autocorrelations would show up for some function g. Second, we can chose functions g such that $Eg^4(X_i) < \infty$, so the existence of the fourth moment is not a problem. We recall that if the X_i are independent, we expect 95% of the ACF values for the $g(X_i)$ to be within the $\pm 2n^{-1/2}$ bands.

We consider the following transformations:

$$g(x) = \ln(x), \quad g(x) = x^{-p}, \ p = 1, 2, 3.$$

Since $X_i \geq 1$, the above transformations are well defined. Direct verifications show that (3.1) implies that $P(\ln X_i > t) \sim ce^{-\alpha t}$, so all moments of $\ln X_i$ are finite. For $p > 0$, the random variables X_i^{-p} values in the interval (0, 1], so to ensure the existence of the fourth moment, we must examine their behavior in neighborhoods of 0. By (3.1),

$$P(X_i^{-p} < x) = P(X_i > x^{-1/p}) \sim cx^{\alpha/p}, \quad \text{as } x \to 0.$$

Therefore, as $x \to 0$, the density is proportional to $x^{\alpha/p-1}$. Hence all moments of X_i^{-p} are finite.

We found that for all 28 links, the ACF plots of $\log(X_i)$, X_i^{-1}, X_i^{-2} and X_i^{-3} show at least 95% correlations within the $\pm 2n^{-1/2}$ bands. Figure 5 shows the ACF plots for the transformed interarrival times at the chin-ipls link. The plots for all links are very similar. The ACF at lag zero is always 1, so it is not plotted.

We have performed a similar analysis using the cross–correlation function (CCF), i.e. we examined the CCF values of $g(X_i(\ell))$ and $g(X_i(\ell'))$, where ℓ and ℓ' are any two links. We have found that at least 95% of these CCF's were within the corresponding confidence bands.

The findings of this section can be summarized as follows:

1. The anomalies interarrival times have approximately Pareto tails; it is reasonable to assume that relation (3.1) holds with $1 < \alpha < 2.5$.
2. It is reasonable to assume the these interarrival times form a sequence of iid random variables.
3. The arrival times in different links form independent sequences.

We emphasize that these findings pertain to the database created by Bandara et al. (2014), but the statistical methodology we presented can be applied to other databases. In the next section, we use these findings to describe the distribution of the waiting time for the arrival of the next anomaly.

4 Distribution of the waiting time for the arrival of the next anomaly

As explained in the introduction, a typical waiting time for the arrival of the next anomaly can be expected to be longer than a typical time between the arrivals of the anomalies. This assertion can be quantified by comparing quantiles of the two distributions. As seen in Table 2, for our database, the difference is quite large. For all links, the quantiles of the waiting time distribution are several times larger than the corresponding quantiles of the distribution of the interarrival times. Intuitively, this is due to the heavy-tailed distribution of the interarrival times established in Sect. 3. Since there are many very long interarrival times, an arbitrary time instant t is likely to be in a long interarrival time, so the the time until the arrival of the next anomaly is more likely to be long than short. We emphasize that unlike the interarrival times, the waiting times are not observed directly. Their distribution must be computed using suitable tools of probability theory. Such tools are provided by the *renewal theory*, which is explained in many textbooks on stochastic processes, e.g. in Kulkarni (2017). Before presenting the computations leading to the estimates in Table 2, we introduce some results of the renewal theory, which are relevant to our task.

Consider a fixed link and recall that $S_1, S_2, \ldots$ are the times of the anomaly arrivals. We have demonstrated in Sect. 3 that the interarrival times $X_i = S_i - S_{i-1}$ are iid random variables with finite expectation (their distribution may very from link to link). These assumptions are enough to apply the results of the renewal theory stated below.

Denote by $\{N(t), t \geq 0\}$ the renewal process with interarrival times X_i, i.e. $N(t)$ is the count of arrivals up to and including time t, i.e. $N(t) = \max\{n : S_n \leq t\}$. Notice that the random time $S_{N(t)+1}$ is the time of the arrival of the next anomaly after time t. Therefore, the waiting time is

$$B(t) = S_{N(t)+1} - t.$$

We will work with the complementary cdf of $B(t)$ defined by

$$H_t(x) = P(B(t) > x), \quad x \geq 0, \ t \geq 0.$$

Using the key renewal theorem, one can show that as t increases, which corresponds, to an asymptotic equilibrium, the probabilities $H_t(x)$ converge to a limit given by

Table 2 Estimated 25th, 50th, 75th, 90th and 95th percentiles of the waiting time distribution (first columns) and the interarrival time distribution (in parentheses)

Link	25th		50th		75th		90th		95th	
atla-hstn	117	(7)	334	(53)	739	(255)	1525	(660)	2209	(1056)
atla-ipls	177	(14)	524	(119)	1398	(420)	2736	(1051)	3615	(1627)
atla-wash	112	(12)	292	(106)	700	(344)	1486	(639)	2052	(1076)
chin-ipls	86	(9)	238	(76)	622	(242)	1896	(479)	2970	(693)
chin-nycm	101	(24)	268	(122)	681	(318)	1623	(623)	2423	(968)
dnvr-kscy	135	(9)	353	(63)	832	(376)	1838	(769)	2443	(1063)
dnvr-snva	76	(6)	213	(37)	563	(185)	1457	(432)	1974	(626)
dnvr-sttl	152	(11)	422	(78)	1027	(354)	1896	(836)	2443	(1652)
hstn-atla	112	(9)	317	(77)	788	(279)	1691	(655)	2306	(1124)
hstn-kscy	130	(7)	375	(84)	949	(295)	1935	(763)	2580	(1234)
hstn-losa	130	(6)	361	(50)	822	(279)	1720	(737)	2384	(1107)
ipls-atla	137	(13)	371	(118)	812	(425)	1437	(964)	1955	(1329)
ipls-chin	81	(12)	217	(83)	495	(250)	1154	(548)	1886	(706)
ipls-kscy	77	(11)	212	(75)	524	(224)	1232	(457)	1789	(722)
kscy-dnvr	82	(12)	223	(84)	524	(246)	1379	(526)	2033	(691)
kscy-hstn	131	(10)	395	(103)	1037	(320)	2130	(749)	2873	(1387)
kscy-ipls	98	(15)	261	(85)	588	(320)	1564	(606)	2228	(767)
losa-hstn	130	(6)	380	(40)	998	(264)	2209	(649)	3009	(1104)
losa-snva	109	(7)	297	(53)	851	(250)	1994	(495)	2658	(940)
nycm-chin	76	(12)	209	(72)	485	(247)	1330	(494)	2013	(632)
nycm-wash	101	(27)	266	(139)	627	(338)	1554	(693)	2306	(893)
snva-dnvr	117	(7)	322	(47)	832	(265)	1955	(568)	2502	(975)
snva-losa	80	(6)	227	(34)	598	(186)	1613	(427)	2170	(603)
snva-sttl	124	(20)	368	(130)	959	(355)	1974	(818)	2599	(1290)
sttl-dnvr	184	(9)	490	(78)	1115	(470)	1955	(1140)	2404	(1852)
sttl-snva	120	(12)	349	(102)	939	(310)	1877	(726)	2384	(1387)
wash-atla	123	(10)	341	(92)	793	(326)	1838	(756)	2755	(1086)
wash-nycm	146	(23)	395	(142)	1008	(438)	2228	(878)	2951	(1443)

$$H^*(x) := \lim_{t\to\infty} H_t(x) = \frac{1}{\tau}\int_x^\infty (1 - G(u))du, \tag{4.1}$$

where

$$\tau = EX_i, \quad G(u) = P(X_i \le u).$$

We note that the parameters τ and $G(\cdot)$ do not depend on i because the X_i are assumed to have the same distribution. The equilibrium cdf of $B(t)$, $1 - H^*(x)$, is therefore given by

$$F_B(x) = 1 - H^*(x) = 1 - \frac{1}{\tau}\int_x^\infty (1 - G(u))du.$$

This leads to the following formula for the density of the waiting time

$$f_B(x) = \frac{d(1 - \tau^{-1} \int_x^\infty (1 - G(u)du)}{dx} = \frac{1}{\tau}(1 - G(x)). \tag{4.2}$$

Denoting suitable estimators by $\hat{\tau}$ and $\widehat{G}(\cdot)$, we can estimate the cdf and the density of the waiting time, respectively, by

$$\widehat{F}_B(x) = 1 - \frac{1}{\hat{\tau}} \int_x^\infty (1 - \widehat{G}(u))du \tag{4.3}$$

and

$$\hat{f}_B(x) = \frac{1}{\hat{\tau}}(1 - \widehat{G}(x)). \tag{4.4}$$

A central issue is therefore 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

$$\widehat{G}(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i \le x\},$$

where N is the count of interarrival times in a given link, and $1\{\cdot\}$ is the set indicator function. The Glivenko–Cantelli theorem asserts that, as $N \to \infty$, $\sup_x |\widehat{G}(x) - G(x)|$ tends to zero with probability 1. The problem of the estimation of the mean, τ in our context, has been given a lot of attention. A textbook estimator is the sample average, but as we demonstrate in Appendix A, this is not a good choice for the X_i we study. Numerical comparisons of various commonly used estimators presented in Appendix A reveals that a very good choice for our data is the estimator which can be derived directly from the empirical cdf $\widehat{G}(\cdot)$ via

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

This estimator is based on the relation $EX = \int_0^\infty P(X > x)dx$, valid for any nonnegative random variable X. It very desirable property is that it lead to an estimator of $f_B(\cdot)$, which is a valid density:

$$\int_0^\infty \widehat{f}_B(x)dx = \frac{1}{\int_0^\infty (1 - \widehat{G}(x))dx} \int_0^\infty (1 - \widehat{G}(x))dx = 1.$$

Using estimator (4.3) we computed the quantiles shown in Table 2. Using (4.4), we computed estimates of the densities of the waiting time. All these densities have qualitatively very similar shapes, so we show just two of them in Fig. 6.

Probabilities of very long waiting times can be computed in an alternative way. Using L'hopital's rule together with the assumptions that $1 - G(X) \approx cx^{-\alpha}$, we

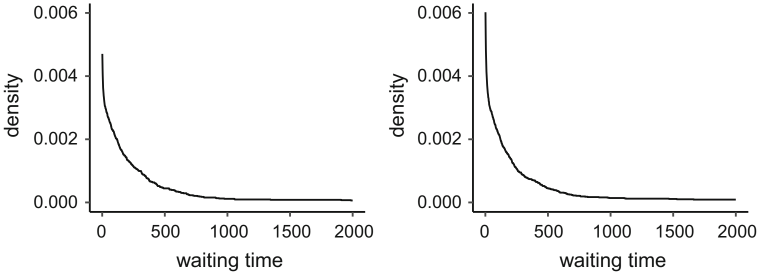

Fig. 6 Estimated densities of the waiting time for the arrival of the next anomaly: left chin-ipls, right dnvr-snva

obtain

$$\lim_{x\to\infty}\frac{1-F_B(x)}{cx^{-(\alpha-1)}/(\tau(\alpha-1))}=\lim_{x\to\infty}\frac{f_B(x)}{cx^{-\alpha}/\tau}=\lim_{x\to\infty}\frac{(1-G(x))/\tau}{cx^{-\alpha}/\tau}=1,$$

which yields

$$1-F_B(x)\approx(\tau(\alpha-1))^{-1}cx^{-(\alpha-1)}. \tag{4.5}$$

For large x, the constant c can be estimated by $x^{\alpha}(1-\widehat{G}(x))$, so (4.5) leads to

$$P(B(t)>x)\approx\frac{1}{(\hat{\alpha}-1)\hat{\tau}x}(1-\widehat{G}(x)).$$

We conclude this section by presenting a simple argument showing that the waiting time distribution is equal to the interarrival distribution if and only if one of them is exponential. Under (3.1) the distribution of the waiting time thus cannot be equal to the distribution of the interarrival time. The postulated equality is equivalent to the equality of the densities, i.e., by (4.2), to

$$\frac{dG(x)}{dx}=\frac{1}{\tau}(1-G(x)).$$

The above linear differential equation has only one solution $G(\cdot)$ which is a cdf; it is

$$G(x)=1-\tau^{-1}e^{-x/\tau},\quad x>0.$$

The following conclusions can be drawn from the research presented in this section, including some details presented in Appendices A and B.

1. For every link, the quantiles of the distribution of the waiting time are generally several times larger then the corresponding quantiles of the distribution of the interarrival times. Very roughly, the interquartile range for the interarrivals is (10, 300), but for the waiting times it is (100, 700) (in units of 5 min). Looking at the medians, a typical time separation between two anomalies is about 8 h, but the

typical waiting time for the arrival of the next anomaly is about 25 h. This paradox is a consequence of the heavy-tailed interarrival times.

2. The waiting time density functions are decreasing. They have large probability masses at both ends, close to zero and close to infinity. Unlike the exponential distribution, they do not touch the vertical line $x = 0$. They decay very slowly as $x \to \infty$, much slower than the exponential distribution.
3. There is a noticeable variation in the distributions of waiting time among the links. Some distributions, e.g. links atla-ipls, sttl-dnvr, wash-nycm, show heavier tails than others, e.g. links dnvr-snva, ipls-kscy, snva-losa. Appendix B presents formal significance testing procedures that confirm that these distributions are not the same for links.
4. Mathematically, the distribution of the waiting time has a heavier tail than the distribution of the interarrival time. Under model (3.1), relation (4.1), implies that in equilibrium, i.e. for large t, $P(B(t) > x) \sim c' x^{-(\alpha-1)}$, as $x \to \infty$. (By (4.5), $c' = (\tau(\alpha - 1))^{-1} c$.) The bounds $1 < \alpha < 2.5$, imply $0 < \alpha - 1 < 1.5$. This means that the random variable describing the waiting time is stochastically so large that it has infinite variance, and may even have infinite expected value (if $\alpha < 2$). In the latter case, for a device placed at a link, the expected waiting time until the arrival of the next anomaly is infinite, but an anomaly will arrive with probability 1.

5 Summary and main conclusions

We studied the arrival times of anomalies in 28 unidirectional links of the Internet2 network. Our objective was to provide information about the distribution of two random variables defined for each of these links: the interarrival time and the waiting time for the arrival of the next anomaly. The statistical characteristics of the interarrival time are accessible directly from physical measurements. The distribution of the waiting time is not directly accessible. It must be estimated using renewal theory.

We have shown that it is reasonable to assume that the interarrival times have heavy-tailed distributions, and they are independently and identically distributed random variables. We derived a parametric model for the distribution of the time between arrivals of two anomalies, see Appendix A, which can be used in various simulations. We have found that the density functions of the waiting time are decreasing with large probability masses for short and very long waiting times. Mathematically, the expected waiting time for the arrival of the next anomaly can be infinite. Practically, we expect the next anomaly to arrive within 6 to 13 days. We have also found that waiting time has heavier tail than interarrival and has infinite variance. There is a significant variation in the waiting time distributions among the 28 links. With formal hypothesis testing, we have confirmed that the interarrival and waiting time distributions among the 28 links are not identical.

We have performed a similar analysis for the *gaps* between the anomalies, which are defined as time separation between the end of the latest anomaly and the beginning of the next anomaly. The gaps are always shorter the the interarrival times, but the differences are very small as the duration of an anomaly is small compared to the large separation times. The modifications thus occur only in the left tail of the distributions, close to zero, and relation (3.1) is unaffected. We conclude that the distribution of these gaps is also heavy–tailed, in particular it is not exponential.

While the conclusions of our statistical and probabilistic analyses pertain to a specific network over a specific time period, it is hoped that the approaches presented in this paper will prove useful in other similar analyses.

Acknowledgements This research has been partially supported by NSF grants DMS–1737795, *DMS 1923142 and CNS 1932413*. We thank Professor Anura P. Jayasumana of the CSU's Department of Electrical and Computer Engineering for sharing the Internet2 anomalies data.

Appendices

A Estimation of the mean time between the arrivals of the anomalies

As explained in Sect. 4, to reliably estimate the distribution of the waiting time, we need a good estimator of the mean interarrival time. In the context of our data, this is a delicate task because the interarrival times have heavy tails, which will bias the usual sample mean. We compared the performance (root mean squared error-RMSE) of the following 12 mean estimators:

Median,
Huber location estimator with varying truncation constants: $k = 5, 10, 20, 30$,
Sample mean,
Trimmed mean with varying trimming fractions: trim = 0.025, 0.05, 0.10, 0.15, 0.20,
The estimator $\hat{\tau} = \int_0^\infty (1 - \widehat{G}(t))dt$, based on the formula $\tau = E(X) = \int_0^\infty P(X > t)dt$.

A crucial question is to generate observation from a distribution which resembles the distribution of the real interarrival times, and whose mean (expected value) can be computed analytically. We can then consider differences between an estimated mean and the true mean within a simulation study. Since interarrival times have many small values and dominating large values with lower frequencies, we use a mixture model: interarrival times come from a Weibull distribution with probability p and from a half-t distribution with probability $1 - p$. The Weibull component is designed to model the occurrence of small values, while the half-t component is designed to model the tail behavior and allow for either finite or infinite variance. (The t distribution satisfies (3.1) with α equal to the degrees of freedom parameter ν.) The density function from which we simulated observations is thus given by

$$f(x) = pf_W(x; k, \lambda) + (1 - p)f_t(x; \nu, \sigma), \quad x > 0,$$

where

$$f_W(x; k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k},$$

$$f_t(x;\nu,\sigma) = 2\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi\sigma^2}}\left[1+\frac{1}{\nu}\frac{x^2}{\sigma^2}\right]^{-(\nu+1)/2}.$$

and where $k, \lambda, \nu, \sigma > 0$ and $0 \leq p \leq 1$. For this model the value of τ can be computed, it is equal to

$$\tau = c\,[\lambda\Gamma(1+1/k)] + (1-c)\left[2\sigma\sqrt{\frac{\nu}{\pi}}\frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)(\nu-1)}\right], \quad \nu > 1.$$

We estimated the mixture model using Maximum Goodness-of-fit Estimator, which minimizes Kolmogorov–Smirnov distance, using the R package *fitdistrplus*. We then used the estimated model to generate $n = 1000$ samples of synthetic interarrival times to compute Monte Carlo RMSE of mean estimators as

$$RMSE = \sqrt{\frac{1}{n}\sum_{r=1}^{n}(\hat{\tau}_r - \tau)^2},$$

where $\hat{\tau}_r$ is a mean estimator computed from the rth Monte Carlo sample, and τ is the mean of the estimated mixture model.

We found that the mixture model has good fit to the observed interarrival times. The fits in all links are similar to those shown in Fig. 2. The Kolmogorov–Smirnov goodness-of-fit test also fails to reject, for all links, the null hypothesis of equal distribution between real and simulated data. In all 28 links, estimates for ν are between 1.2 and 2.2, so the half-t distribution successfully captures the tail behavior inferred from the Hill plots. Since the ν estimates are all greater than 1, the means of estimated mixture distributions exist.

The RMSEs for the sample mean, the most commonly used estimator, and the three best estimators are shown in Table 3. We see that sample mean performs poorly. The estimator $\hat{\tau}$, which we used in Sect. 4, is most often the best estimator, and when it is not, its RMSE is very close to the lowest RMSE. This justifies its choice as the preferred mean estimator for the interarrival time.

B Significance tests

We present here formal statistical significance tests that confirm the conclusions stated in Sect. 4. We first consider the testing problem:

H_0 : The distributions of interarrival times are identical for the 28 links,
H_A : The distributions of interarrival times are **not** identical for the 28 links.

Since, as shown in Sect. 4, $f_B(x) = \tau^{-1}(1 - G(x))$, this test also applies to the distributions of waiting times. If these distributions are equal, then their expected values are also equal. We therefore use a permutation test based on the usual F-

Table 3 RMSEs of the sample mean and three best mean estimators for interarrival time; bold indicates the lowest RMSE among the 12 estimators we considered

	Sample mean	Huber $k = 20$	Huber $k = 30$	$\hat{\tau}$
atla-hstn	208.41	191.77	174.83	**139.43**
atla-ipls	740.02	248.86	223.40	**179.47**
atla-wash	128.35	48.18	43.21	**39.90**
chin-ipls	20.25	**15.80**	16.09	17.53
chin-nycm	36.02	**26.00**	26.12	27.63
dnvr-kscy	103.53	71.93	61.74	**50.02**
dnvr-snva	14.51	13.01	**11.92**	12.55
dnvr-sttl	144.57	82.83	72.76	**61.73**
hstn-atla	48.52	35.97	32.26	**31.13**
hstn-kscy	125.96	62.98	54.48	**46.01**
hstn-losa	205.54	124.27	108.75	**70.71**
ipls-atla	90.86	68.18	61.75	**57.26**
ipls-chin	112.56	64.67	56.98	**46.15**
ipls-kscy	23.32	19.07	**17.95**	18.68
kscy-dnvr	318.20	51.64	45.14	**37.70**
kscy-hstn	47.05	33.33	**31.86**	32.90
kscy-ipls	86.54	32.17	**29.80**	29.87
losa-hstn	101.31	111.72	95.64	**59.51**
losa-snva	36.73	29.83	26.43	**26.03**
nycm-chin	362.90	238.45	223.98	**190.55**
nycm-wash	45.83	**30.99**	31.08	32.38
snva-dnvr	55.94	32.80	28.81	**27.31**
snva-losa	22.43	15.38	**13.55**	13.87
snva-sttl	77.55	50.74	47.28	**46.21**
sttl-dnvr	338.27	172.03	152.64	**119.24**
sttl-snva	35.93	**23.81**	24.13	26.14
wash-atla	122.29	68.02	58.87	**50.96**
wash-nycm	340.50	143.80	129.78	**119.12**

statistic:

$$F = \frac{U}{V}, \quad U := \frac{\sum_{i=1}^{28} (\overline{X}_{i.} - \overline{X}_{..})^2 n_i}{28 - 1}, \quad V := \frac{\sum_{i=1}^{28} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_{..})^2}{N - 28}, \qquad \text{(B.1)}$$

where $\overline{X}_{i.}$ is the sample mean of interarrival times in link i, $\overline{X}_{..}$ is the sample mean of interarrival times across all links, n_i is the number of observed interarrival times for link i, N is the number of observed interarrival times in all 28 links.

The observed value of the test statistic is $F = 5.52$. However, we cannot compare it to a tabulated critical value because the distribution of the interarrival times is not normal. We therefore estimate the null distribution using permutations, see e.g. Good

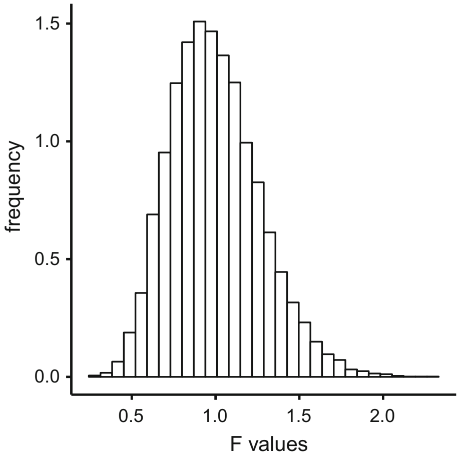

Fig. 7 Sampling null distribution of the F statistic (B.1) based on ten thousand permutations

(2013). Under H_0, the interarrival times among the 28 links are iid random variables; hence, by randomly reassigning the N interarrival times to the 28 groups, such that the number of observations in each group is not changed, we produce a new pseudo dataset for which H_0 is true. We resample this way for 10,000 times, and obtain the null distribution of the test statistics shown in Fig. 7. It is seen that the observed value of $F = 5.52$ is far to the right of the range of the test statistics under the null hypothesis. Formally, we approximate the p-value with the proportion of samples with $F > 5.52$, and see that $p - value < 0.0001$. As the result, we reject H_0.

We also performed the Anderson–Darling test with the R package *kSamples*. The standardized Anderson–Darling test statistics is 28.36 with $p - value < 0.0001$. Hence, we also reject H_0.

We conclude that the interarrival time distributions among the 28 links are not identical; hence the waiting time distributions among the 28 links are not identical either.

We also used three standard goodness-of-fit tests implemented with R package *EnvStats*: Kolmogorov–Smirnov test, Cramer-von Mises test and Anderson–Darling test to check if the distribution of the interarrival times is exponential. For all 28 links, and for each test, the null hypothesis of an exponential distribution is rejected at the significance level of 5 percent. We conclude that the anomaly interarrival time does not have an exponential distribution.

References

Adler R, Feldman R, Taqqu MS (1998) A practical guide to heavy tails: statistical techniques for analyzing heavy tailed distributions. Birkhauser, Boston

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

Bhuyan MH, Bhattacharyya DK, Kalita JK (2014) Network anomaly detection: methods, systems and tools. IEEE Commun Surv Tutor 16:303–336

Chandolla V, Benerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(15):58

Good PI (2013) Permutation, parametric, and bootstrap tests of hypotheses. Springer, Berlin

Hall P (1990) Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. J Multivar Anal 32:177–203

Kallitsis M, Stoev S, Bhattacharya S, Michailidis G (2016) AMON: an open source architecture for online monitoring, statistical analysis and forensics of multi-gigabit streams. IEEE J Sel Areas Commun 34:1834–1848

Kulkarni VG (2017) Modeling and analysis of stochastic systems. Chapman and Hall, Atlanta

Leland WE, Taqqu MS, Willinger W, Wilson DV (1994) On the self-similar nature of ethernet traffic (extended version). IEEE/ACM Trans Netw 2:1–15

Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y (2013) Intrusion detection system: a comprehensive review. J Netw Comput Appl 36:16–24

Park K, Willinger W (2000) Self-similar network traffic and performance evaluation. Wiley, Hoboken

Paschalidis IC, Smaragdakis G (2009) Spatio-temporal network anomaly detection by assessing deviations of empirical measures. IEEE/ACM Trans Netw 17:685–697

Peng L, Qi Y (2017) Inference for heavy-tailed data analysis: applications in insurance and finance. Academic Press, Cambridge

Resnick SI (1997) Heavy tail modeling and teletraffic data. Ann Stat 25:1805–1869

Resnick SI (2007) Heavy-tail phenomena: probabilistic and statistical modeling. Springer, Berlin

Shumway RH, Stoffer DS (2017) Time series analysis and its applications with R examples. Springer, Berlin

Tsai C-F, Hsu Y-F, Lin C-Y, Lin W-Y (2009) Intrusion detection by machine learning: a review. Expert Syst Appl 39:11994–12000

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

Xie M, Han S, Tian B, Parvin S (2011) Anomaly detection in wireless sensor networks: a survey. J Netw Comput Appl 34:1302–1325

Zarpelao BB, Miani RS, Kawakani CT, de Alvarenga SC (2017) A survey of intrusion detection in internet of things. J Netw Comput Appl 84:25–37