2

Statistical Inspection and Analysis Techniques for Traffic Data Arising from the Internet

Natalia M. Markovich[1] and Udo R. Krieger[2]

[1]*Institute of Control Sciences, Russian Academy of Sciences, Profsoyuznaya str. 65, 117997 Moscow, Russia; e-mail: markovic@ipu.rssi.ru*
[2]*Information Systems and Applied Computer Science Department, Otto-Friedrich University, D-96045 Bamberg, Germany; e-mail: udo.krieger@ieee.org*

Abstract

The statistical characterization of measurements arising from Internet traffic requires specific analysis and estimation techniques since the underlying traffic characteristics are often distributed with heavy tails. Here we provide a survey of fast and simple methods to detect the heaviness of tails and dependence in traffic data. We consider the dependence of both univariate and bivariate data. Important practical problems regarding the dependence between the transmission rate and file size or the time of transmission are discussed as well. Finally, the recommended tools are illustrated by applications to real Web, TCP and video data.

Keywords: Traffic characterization, heavy tails, data dependence, extremal index, Pickands' function.

Abbreviations

ACF	autocorrelation function
a.s.	almost sure

D. D. Kouvatsos (ed.), Traffic and Performance Engineering for Heterogeneous Networks, 41–60.

DF	distribution function
EVI	extreme value index
HTML	hypertext markup language
i.i.d.	independent, identically distributed
LRD	long range dependence
PDF	probability density function
r.v.	random variable
TCP	transmission control protocol

2.1 Introduction

Currently, traffic measurements of high-speed packet-switched networks that are taken with high resolution at different time scales constitute an important topic of data mining and teletraffic engineering in the Internet. Associated issues such as the statistical data analysis of the measurements at the session and flow levels, the on-line estimation of the underlying random characteristics of the developed Web and Internet traffic models and network management actions based on on-line traffic analysis are studied intensively, too. In this regard it is necessary to analyze the traffic characteristics of the gathered data in a rigorous mathematical manner and to cope very carefully with the related tasks of model selection, estimation and assessment.

Considering the transported traffic in IP networks, besides the dominant peer-to-peer traffic a large portion of the flows are currently still generated by classical client-server applications of the Web (cf. [7, 15, 16]). From the statistical perspective, many observed Web traffic characteristics are determined by independent random variables and often distributed with heavy-tails (see [4, 14, 24]).

Examples of independent univariate data are given by file sizes and session volumes transmitted to an individual customer. Inter-arrival times of packets in aggregated traffic may be dependent. The pairs (file size, duration of its transmission) generated by an individual client-server relationship are independent, but components of these pairs are evidently dependent. Generally, the dependence is not always obvious and requires testing.

Considering the data analysis, the specific features of heavy-tailed distributions are the following:

(1) the tail of the distribution tends to zero at infinity slower than an exponential tail;

(2) the possible non-existence of all or some higher moments;
(3) sparse data at the tail domain of the distribution.

Examples of heavy-tailed distributions are provided by Pareto, Lognormal, and Cauchy distributions as well as Weibull distributions with shape parameter less than 1. They require specific statistical methods for the rigorous investigation of the data (see [18]).

The reconstruction of such heavy-tailed distributions is a difficult issue. The histogram, for instance, provides a good estimate of the controlling probability density function (PDF). In the case of distributions with heavy tails, however, it shows a misleading estimate in the tail domain or over-smoothes the main part of the PDF. The same is true for kernel estimators (cf. [28]). To improve the estimation of a heavy-tailed PDF at the tail, a preliminary transformation of the data from a Pareto to a triangular distribution has been used in [18]. Kernel estimates with variable bandwidth kernels provide another way to reconstruct such PDFs (cf. [28]).

In practice important traffic characteristics like transfer delays often require the analysis of quantiles of the corresponding distribution. Usually, quantiles can be estimated by means of an empirical distribution function (DF). However, high quantiles like 99%, 99.9%, cannot be calculated in such a way since the empirical DF is equal to one outside the range of the empirical sample. In [21] a new estimate for high quantiles has been proposed and its asymptotic normality was proved in [17].

These examples show that it is important to recognize the heavy tails in the traffic characteristics before their further analysis. To resolve it, formal nonparametric tests (cf. [12, 25]) or rough statistical methods for heavy-tailed features can be applied (cf. [6, 18]).

Considering the analysis of heavy-tailed data, the extremes, e.g., large file sizes, long durations of transmission etc., influence significantly on the rate of transmission (see [10, 20]). In this situation the asymptotic distribution (as the sample size tends to infinity) of the maximum of the sample called extreme value distribution is used as a model of the DFs determining the extremes of traffic characteristics (see [18]). Such models include a parameter called tail index. The tail index $\alpha \in \mathbb{R}$ (or the extreme value index (EVI) $\gamma = 1/\alpha$) shows the shape of the tail.

In this context, one has to test the dependence in the underlying univariate data since statistical methods usually require independence and stationarity of the distribution. Often, it is also interesting to know the dependence between pairs of traffic characteristics.

For this purpose we present in this paper several procedures that may help us to detect heavy tails and to analyze the dependence structure of a gathered sample with Internet data.

The material is organized as follows. In Section 2.2 we sketch simple inspection techniques for heavy-tailed traffic characteristics. We discuss statistical procedures to test the dependence in univariate data in Section 2.3 and to determine the dependence in bivariate data in Section 2.4. Finally, we present some conclusions.

2.2 Testing the Heaviness of Tails

2.2.1 Definitions

The formal definitions of heavy-tailed distributions, their subclasses and theoretical properties can be found in [1, 6, 18]. Here we only define the most important and widest class of regularly varying heavy-tailed distributions and formulate its important property.

Definition 1. *The class $\Re_\alpha$ of distributions with regularly varying tails and the tail index $\alpha = 1/\gamma$, $\gamma > 0$, is defined by*

$$\Re_\alpha = \{F : \mathbb{R} \to [0,1] \mid 1 - F(x) = x^{-\alpha}\ell(x), 0 < x, 0 < \alpha \in \mathbb{R}\},$$

where $\ell(x)$ is a slowly varying function that satisfies the condition $\lim_{x\to\infty} \ell(tx)/\ell(x) = 1$ for all $t > 0$.

Definition 2. *In the case $\alpha = 0$ $\Re_\alpha$ determines a subclass of distributions called super heavy-tailed. These distributions have no finite moments of any order.*

The theoretical property of distributions with regularly varying tails that is important for practice concerns the finiteness of moments. More exactly, the pth moment of the random variable (r.v.) X_1 exists, i.e., $E|X_1|^p < \infty$ holds, if the tail index α satisfies $0 < p < \alpha$ and the distribution belongs to class $\Re_\alpha$ (cf. [6, p. 330]).

The detection of heavy tails and super heavy tails may be provided by formal tests, see [12] and [25], respectively.

2.2.2 Rough Preliminary Methods

There are some simple statistical procedures that allow us to detect heavy tails. These include the ratio of the maximum to the sum, the plot of the empirical mean excess function, the QQ-plot and the tail index estimation.

Let $X_1, \dots, X_n$ be i.i.d. r.v.s. The statistic (cf. [6, p. 308]) $R_n(p) = M_n(p)/S_n(p)$, $n \geq 1$, $p > 0$, where $S_n(p) = |X_1|^p + \cdots + |X_n|^p$, $M_n(p) = \max(|X_1|^p, \dots, |X_n|^p)$, $n \geq 1$ indicates the amount of finite moments. More exactly, if $R_n(p)$ is small for large n, then $E|X|^p < \infty$, otherwise it suggests that the pth moment is infinite, i.e. $E|X|^p = \infty$. This suggestion is based on the theoretical property

$$R_n(p) \xrightarrow{\text{a.s.}} 0 \qquad \Leftrightarrow \qquad E|X|^p < \infty.$$

It is the idea of a QQ-plot ("quantiles against quantiles"-plot) to draw the dependence $\{(X_{(k)}, F^{\leftarrow}((n-k+1)/(n+1))) : k = 1, \dots, n\}$, where $X_{(1)} \geq \dots \geq X_{(n)}$ are the order statistics of the sample, and $F^{\leftarrow}$ is an inverse function of the DF F. Usually, the QQ-plot is built as a dependence of exponential quantiles against the order statistics of the underlying sample. Generally, one can select any distribution instead of an exponential one. The QQ-plot looks close to linear if the model of the distribution F is selected properly.

The mean excess function $e(u) = E(X - u | X > u)$, $0 \leq u < X_F \leq \infty$, where $X_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ is the finite right endpoint of the distribution, can be tested in practice by its empirical analogue

$$e_n(u) = \sum_{i=1}^{n} (X_i - u)\mathbf{1}\{X_i > u\} \Big/ \sum_{i=1}^{n} \mathbf{1}\{X_i > u\}.$$

Here $\mathbf{1}\{A\}$ denotes the indicator of an event A. The plot of the mean excess function tends to infinity for heavy-tailed distributions (e.g., for a Pareto distribution it is linear), decreases to zero for light-tailed distributions and remains constant for an exponential distribution.

The tail index or its inversion EVI is the most important characteristic of a heavy-tailed distribution. There are numerous procedures to estimate it. The most popular estimators are given by the Hill's estimator [11]

$$\widehat{\gamma}^H(n,k) = \frac{1}{k} \sum_{i=1}^{k} \log X_{(n-i+1)} - \log X_{(n-k)},$$

which is used to estimate the positive EVI γ of a heavy-tailed r.v. X, and the moment estimator [1]

$$\hat{\gamma}^M(n,k) = \hat{\gamma}^H(n,k) + 1 - 0.5\left(1 - (\hat{\gamma}^H(n,k))^2/S_{n,k}\right)^{-1}, \qquad (2.1)$$

where $S_{n,k} = (1/k)\sum_{i=1}^{k}(\log X_{(n-i+1)} - \log X_{(n-k)})^2$. $\hat{\gamma}^M(n,k)$ is also valid for real valued EVIs, although it has a larger asymptotic variance than $\hat{\gamma}^H(n,k)$. For both estimators the parameter k indicates the largest order statistics $X_{(n-k)} \le \ldots \le X_{(n)}$ of the underlying sample $X_1, \ldots, X_n$ of size n.

Here we present a rather new estimator proposed in [5] that is only valid for positive EVIs. In [18] it has been called the group estimator. Its advantage is determined by the recursive property that gives rise to use it for on-line calculations. According to this estimator the sample of i.i.d. r.v.s. $X_1, \ldots, X_n$ is divided into l groups $V_1, \ldots, V_l$, each group containing m r.v.s., i.e. $n = l \cdot m$. Let

$$M_{li}^{(1)} = \max\{X_j : X_j \in V_i\}$$

and let $M_{li}^{(2)}$ denote the second largest element in the same group V_i. Then the statistic

$$\gamma_l = 1/z_l - 1, \qquad z_l = (1/l)\sum_{i=1}^{l} M_{li}^{(2)}/M_{li}^{(1)}$$

is suggested as an estimate of γ.

All estimators of the tail index are very sensitive to the choice of the smoothing parameters. The latter are determined by k in the cases of the Hill's and moment estimators and the amount of observations m in each group in case of the group estimator. The use of plots where the estimator is represented against a smoothing parameter is the easiest way to select these parameters. One can plot $\{(m, z_m), m_0 < m < M_0\}$, $m_0 > 2$, $M_0 < n/2$ or draw a Hill plot $\{(k, \hat{\gamma}^H(n,k)), 1 \le k \le n-1\}$ and then choose the estimate of z_m or $\hat{\gamma}^H(n,k)$ from an interval in which these functions demonstrate stability.

2.2.3 Application to Web Data

In this subsection we present an example of the application of EVI estimators to Web data. This data have been gathered in the Ethernet segment of the Department of Computer Science at the University of Würzburg and have been analyzed in [14, 18, 21]. The data comprise superimposed traffic flows

Table 2.1 Estimation of the extreme value index for Web traffic characteristics.

Estimate	s.s.s.	d.s.s.	s.r.	i.r.t.
$\hat{\gamma}^H(n,k)$	0.92	0.7	0.92	0.65
γ_l	0.84	0.7	0.8	0.5
$\hat{\gamma}^M(n,k)$	0.8	0.7	0.92	0.6

of client-server sessions monitored at the client site and contain basic characteristics of sub-sessions, i.e., the size of a sub-session (s.s.s) in bytes and its duration (d.s.s.) in seconds, as well as the characteristics of the transferred Web pages, i.e., the size of the response (s.r.) in bytes and the inter-response time (i.r.t.) in seconds. The sample sizes n of both d.s.s. and s.s.s. are given by 373 whereas $n = 7107$ is used for both i.r.t. and s.r.. A detailed description of these data can be found in [18].

In Table 2.1 and Figure 2.1 one can see the estimation of the EVI for the data samples s.s.s., d.s.s., i.r.t. and s.r. by means of the Hill's, group and moment estimators. The parameter m of the group estimate γ_l and the parameter k of both the Hill's and moment estimates were selected by the plots of these estimates against the corresponding parameters. The straight horizontal lines in the plots correspond to the intervals of stability of the EVI estimates and, therefore, state the estimated values of the EVI. Indeed, one may select several stability intervals. Regarding s.s.s., for example, the value 0.69 corresponds to the first stability interval of the moment estimate. Regarding d.s.s. one could also select 0.6 as appropriate Hill's and moment estimates. This feature demonstrates the obvious disadvantage of the selection by a plot and the necessity to use other data-driven methods. The bootstrap procedure is such an alternative method (see [1, 18]).

Observing the estimates of γ, one may conclude that the estimates of the tail index $\alpha = 1/\gamma$ are all positive. It implies that the distributions of all considered Web characteristics are heavy-tailed. Moreover, the α's are less than 2 for all considered data sets. It follows from extreme value theory [6] that the βth moments, $\beta \geq 2$, of the distributions of s.s.s., s.r., d.s.s. and i.r.t. are not finite. This feature follows subject to the assumption that the distributions of all characteristics are regularly varying. The tails of the s.s.s. and s.r. distributions are heavier than the tails of d.s.s. and i.r.t. since their EVIs γ are larger.

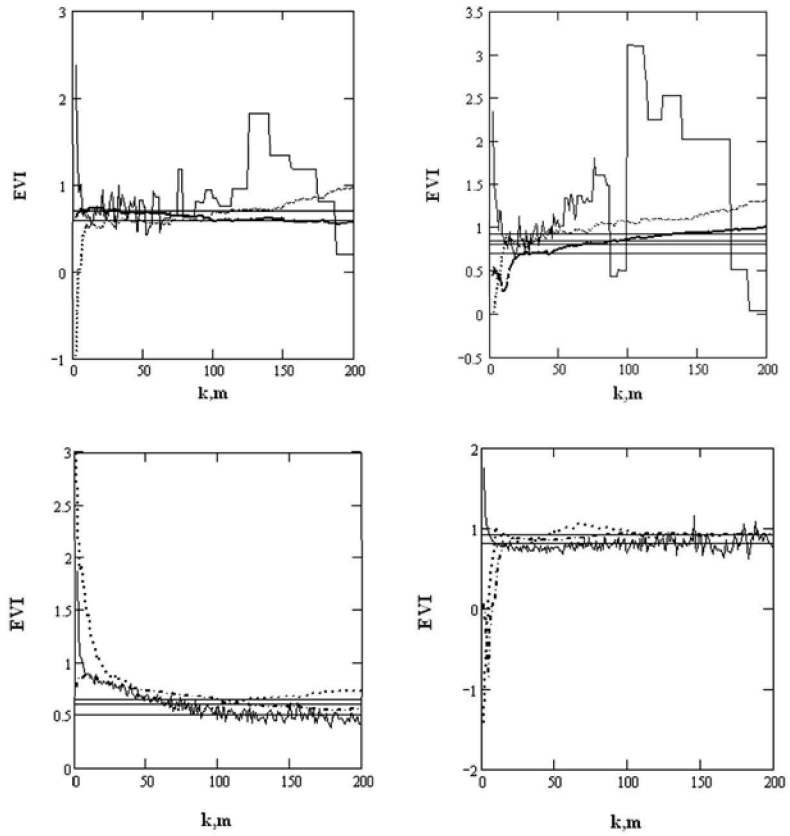

Figure 2.1 EVI estimation by the Hill's estimator (dotted line), moment estimator (dashed line) and the group estimator (thin solid line) for the data sets d.s.s., s.s.s., i.r.t., s.r. (from left to right).

2.3 Testing the Dependence of Univariate Data

2.3.1 Classical Measures

The methods stated in Section 2.2 mostly assume that the measurements are independent. Testing independence constitutes a significant part of the analysis. Unfortunately, it is impossible in practice to check formal independence conditions like, e.g., a strong mixing condition, since it requires the knowledge of the marginal and bivariate distributions. Hence, the autocorrelation function (ACF)

$$\rho_X(h) = \rho(X_t, X_{t+h}) = E\left((X_t - EX_t)(X_{t+h} - EX_{t+h})\right) / \mathrm{Var}(X_t)$$

is considered as an important indicator of the dependence structure of a time series.

The standard sample autocorrelation function at lag $h \in \mathbb{Z}$ is determined by

$$\rho_{n,X}(h) = \frac{\sum_{t=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n)}{\sum_{t=1}^{n}(X_t - \overline{X}_n)^2}, \qquad \overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t. \tag{2.2}$$

The naive analysis of the ACF includes the comparison of values of the ACF at different lags h and the Bartlett's bounds $\pm 1.96/\sqrt{n}$ (see [2]). Then the observations are assumed to be independent with probability 95% if the ACF falls inside the Bartlett's interval. The Bartlett's bounds are valid only for linear processes with Gaussian noise. The use of this convenient interval is meaningless if the underlying r.v.s. are not Gaussian or a time series belongs to a non-linear process. Moreover, the bounds for heavy-tailed regularly varying distributions that are typical for telecommunications have a complicated form (see [27, §9.5]).

$\rho_{n,X}(h)$ may be inaccurate if the sample size n is small or h is close to n. The relevance of this estimate is determined by its rate of convergence to the real ACF. If the variance is infinite (we have observed this case for Web data in Section 2.2.3), $\rho_{n,X}(h)$ cannot be calculated. In [27, p. 342] it was therefore proposed to use the modified ACF estimate

$$\widetilde{\rho}_{n,X}(h) = \frac{\sum_{t=1}^{n-h} X_t X_{t+h}}{\sum_{t=1}^{n} X_t^2} \tag{2.3}$$

without the centering by the sample mean. This estimate may not be relevant for non-linear processes. In this context the selection of a proper type of process is a complicated problem.

Hence, we see that the conclusions regarding independence by the observation of the ACF may be unreliable. One can just assume that the observations may be independent if the values of the ACF are small at large lags and fall inside the Bartlett's interval.

Sometimes, the dependence in the time series $\{X_t, t \geq 0\}$ is long-range, i.e. it remains significant over a long time interval. This implies that even if the values of the ACF are small for large lags their cumulative sum may be large, namely

$$\sum_{h=0}^{\infty} |\rho_X(h)| = \infty. \tag{2.4}$$

Let the ACF be represented for some constant $c_\rho > 0$ by the model

$$\rho_X(h) \sim c_\rho h^{2(H-1)} \quad \text{for large } h,$$

that satisfies the condition (2.4) for $H \in (0.5, 1)$. Then the closer the Hurst parameter H is to one, the deeper is the possible long-range dependence. The Hurst parameter can be calculated by a formula

$$\hat{H}_n = 0.5\left(1 + \log_2(1 + \rho_{n,X}(1))\right)$$

proposed in [13]. It is obtained under the assumption that the process X_t is second-order self similar with the ACF (see [18, p. 47])

$$\rho_X(h) = \frac{1}{2}\left(|h+1|^{2H} - 2|h|^{2H} + |h-1|^{2H}\right).$$

2.3.2 Estimation of the Extremal Index

Instead of the ACF, we consider here the extremal index as an alternative dependence measure. It arises from the theory of extreme values.

Let $X_1, X_2, \ldots, X_n$ be n (not necessarily independent) r.v.s. arising from some stationary process with marginal DF $F(x)$. Denote by $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$ an associated independent sequence with the same DF. Then it follows for large n and u_n that

$$P\{\max(X_1, \ldots, X_n) \le u_n\} \approx P^\theta\left\{\max(\widetilde{X}_1, \ldots, \widetilde{X}_n) \le u_n\right\} = F^{n\theta}(u_n). \tag{2.5}$$

Here, $\theta \in [0, 1]$ is a constant called the extremal index (see [1, 29]). It is clear that $\theta = 1$ holds for i.i.d. sequences. Formula (2.5) implies that the extremal index shows the change in the limiting distribution of the maxima of the sample due to the dependence.

The extremal index has a deeper meaning than only the indication of the dependence. First, it characterizes a cluster structure in the data. The clustering can also be visible on the ACF plots, which implies dependence in the data. For example, the extremal index allows us to divide video data into classes according to the different dependence of frames in the scenes (see [22]).

If, on the other hand, exceedances $\{X_i - u\}$, $i = 1, \ldots, n$ over a threshold u are considered, θ characterizes the distribution of the inter-exceedance times. A representation of the exponential distribution of the inter-exceedance times with an intensity equal to θ was proved in [8].

Among several estimators of the extremal index we want to mention here the blocks and runs estimators. These are only distinguished by the definition of the cluster. In a way, the extremal index is close to the mean excess function. But the latter shows the sample mean excess over the whole sample and the extremal index over the cluster.

One can define a cluster as a block of the data with at least one exceedance over the threshold. Then the blocks estimator is calculated by the formula:

$$\overline{\theta}^B(u) = \frac{k^{-1}\sum_{j=1}^{k}\mathbf{1}(M_{(j-1)r,jr} > u)}{rn^{-1}\sum_{i=1}^{n}\mathbf{1}(X_i > u)}, \tag{2.6}$$

where $M_{i,j} = \max(X_{i+1}, \ldots, X_j)$, k is the number of blocks, $r = [n/k]$ is the number of observations in each block, and $[\cdot]$ denotes the integer part of the number. $1/\overline{\theta}^B(u)$ can be simply interpreted as the ratio of the number of observations that exceed the threshold u to the number of clusters. It shows the mean number of exceedances in a cluster.

If we define a cluster as a block of data with some number of exceedances over the threshold and at the same time the r subsequent observations are all below the threshold u, we get the runs estimator:

$$\overline{\theta}^R(u) = \frac{(n-r)^{-1}\sum_{i=1}^{n-r}\mathbf{1}(X_i > u, M_{i,i+r} \le u)}{n^{-1}\sum_{i=1}^{n}\mathbf{1}(X_i > u)} \tag{2.7}$$

In this case no block structure is required. The runs estimate has a better asymptotic bias than the blocks estimate (see [29]).

The selection of the smoothing parameters k in (2.6) or r in (2.7) constitutes a common problem of both estimators. Data-driven methods of this selection pose an open problem. It is the underlying idea to select those parameters which make the clusters independent. The theory is stated in [6, sect. 8.1]. Very roughly speaking, clusters should be sufficiently far away from each other.

The simplest way is to estimate a θ that corresponds to the stable interval of the plot $(1/\overline{\theta}(u), u)$ over a range of thresholds for a fixed parameter k (or r). The reason is that both considered estimates are consistent, i.e. $\theta = \lim_{n\to\infty}\overline{\theta}$.

Sometimes, it is possible to select these parameters by the nature of the problem. In [22], it has been proposed to take scenes of video data as blocks. This selection was motivated by the scene changes with large variations in the bit rate among different scenes. This feature is typical for video traffic due to the visual shifting between scenes. Hence, such blocks have no equal

sizes. In this respect the blocks estimator has been modified to a scene blocks estimator

$$\overline{\theta}_S^B(u) = \frac{\sum_{j=1}^{k} \mathbf{1}(M_{\sum_{m=0}^{j-1} r_m, \sum_{m=1}^{j} r_m} > u)}{\sum_{i=1}^{n} \mathbf{1}(X_i > u)},$$

where r_j is the number of frames in the jth scene, $\sum_{j=1}^{k} r_j = n$, $r_0 = 0$ and k is the number of scenes.

2.4 Testing the Dependence of Bivariate Data

2.4.1 Dependence between the Rate, File Size, and Duration of TCP Traffic Flows

In teletraffic engineering we often need to detect the dependence of two r.v.s., such as the basic characteristics of TCP flows. The joint behavior of large values of the file size S and the duration D of a transfer and the throughput or rate R of a session, that is $R = S/D$, is considered by many authors due to its practical importance (cf. [10, 20, 23, 26, 27]).

Regarding this investigation one can summarize the following problems:

- There are ties in the data of the flow size and, hence, the distribution of the size is not continuous.
- Physical limitations in the sizes, durations and rates occur due to differences in the access links or due to the impact of the TCP congestion window and self-congestion.
- No simple classical models for the distributions of size, duration and rate arise, but more complicated structures such as mixtures of Lognormal or Pareto distributions, super heavy-tailed Log-Pareto distributions or regularly varying distributions with varying tail indexes.
- A non-homogeneous dependence structure appears.

In [26] truncated univariate and bivariate Lognormal distributions have been proposed to model the flow size, duration and rate distributions. $\log R = \log(S/D) = \log S - \log D$ is investigated instead of S/D. Further the applied regression models $E(\log S | \log D = y)$ and $E(\log D | \log S = x)$ have been built under the assumption that $\log D$ and $\log S$ are governed by truncated univariate normal distributions. Fortunately, the truncation of the distribution does not reflect much on the regression analysis.

In [27, p. 239] the size, duration and rate are assumed to follow regularly varying distributions with tail indices α_S, α_D and α_R and the independence of

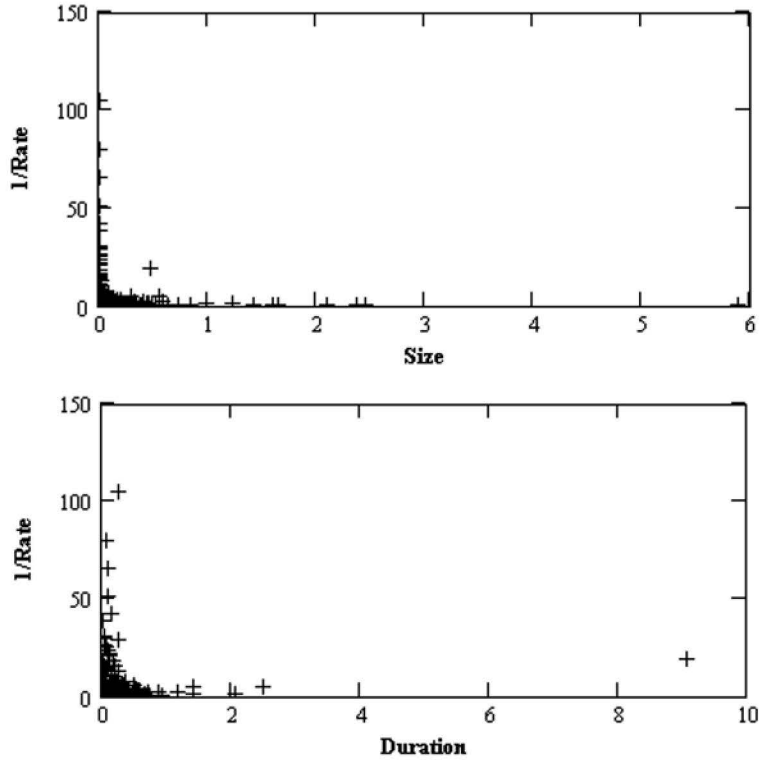

Figure 2.2 Scatter plots of inverse Web session throughput against the size of sub-sessions (top) and duration of sub-sessions (bottom). The "axis hugging" is visible on the top plot suggesting the independence of large sizes and throughputs. For large values the independence of durations and throughputs is not so strict.

R and D is shown by a comparison of the tail indices of the size, α_S, and of $D \cdot R$, $\alpha_D \alpha_R / (\alpha_D + \alpha_R)$, based on Breiman's theorems.

In [10] the independence of R and S and some dependence between R and D are illustrated using data of HTTP responses. The authors have recognized that different strengths of dependence arise between large values and moderate values of a bivariate vector, i.e. the probability of both r.v.s. being large is negligible in case of asymptotic independence. As a consequence, the "axis hugging" by data points at a scatter plot implies extremal independence. The same conclusions regarding the dependence of the pairs (R, S) and (R, D) follow from Figure 2.2. It shows scatter plots of the Web data described in Section 2.2.3.

In [23] the dependence of (D, R) is shown for 80% of the investigated aggregated flows by the examination of the relation $ED \cdot ER/ES = 1$.

In [20] the weak dependence of (S, R) and almost independence of (D, R) of the analyzed TCP-flow data is revealed by the examination of a Pickands' dependence A-function. Subsequently, we shall consider this approach in more detail.

2.4.2 Testing the Dependence by Pickands' Function

The Pickands' function stems from the representation of the limiting distribution of bivariate maxima. It is a convenient form to detect the dependence of extremes of two underlying r.v.s. More exactly, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a bivariate i.i.d. sample with a bivariate extreme value distribution G. It implies that for some normalizing constants $0 < a_{j,n} \in \mathbb{R}$ and $b_{j,n} \in \mathbb{R}$, $j = 1, 2$,

$$P\{\left(M_{1,n} - b_{1,n}\right)/a_{1,n} \leq x, \left(M_{2,n} - b_{2,n}\right)/a_{2,n} \leq y\} \to G(x, y), \quad n \to \infty$$

holds, where $M_{1,n} = \max\{X_1, \ldots, X_n\}$, $M_{2,n} = \max\{Y_1, \ldots, Y_n\}$ are the component-wise maxima.

We consider the representation

$$G(x, y) = \exp\left(\log\left(G_1(x)G_2(y)\right) A\left(\frac{\log G_2(y)}{\log\left(G_1(x)G_2(y)\right)}\right)\right),$$

where A is Pickands' function, G_1 and G_2 are univariate extreme value DFs of the maxima $M_{1,n}$ and $M_{2,n}$, i.e. they are limiting DFs of these maxima themselves. In general, the vector $(M_{1,n}, M_{2,n})$ will not be present in the original data.

Let the random pair (X^*, Y^*) have the DF $G(x, y)$. In the bivariate case, the function $A(t)$ satisfies $A(0) = A(1) = 1$, it is convex and lies inside the triangle determined by the points $(0, 1)$, $(1, 1)$ and $(0.5, 0.5)$. The case $A \equiv 1$ corresponds to total independence and the case $A(t) = \max\{(1-t), t\}$ to total dependence of X^* and Y^*.

The correlation between X^* and Y^*

$$\rho = \int_0^1 \frac{dw}{(A(w))^2} - 1$$

is always nonnegative since $A(w) \leq 1$ holds (cf. [30]).

In practice, the A-function is evaluated by component-wise maxima over m blocks of data (X_j^*, Y_j^*), $j = 1, \ldots, m$. The best known estimators of the A-function include

$$\widehat{A}_m^{HT}(t) = \left((1/m) \sum_{j=1}^{m} \min \left(\frac{\xi_j / \overline{\xi}_m}{1-t} \frac{\eta_j / \overline{\eta}_m}{t} \right) \right)^{-1}$$

proposed in [9], and

$$\log \widehat{A}_m^{C}(t) = \frac{1}{m} \sum_{j=1}^{m} \log \max \left(t\xi_j, (1-t)\eta_j \right) - t \frac{1}{m} \sum_{j=1}^{m} \log \xi_j - (1-t) \frac{1}{m} \sum_{j=1}^{m} \log \eta_j$$

proposed in [3]. Both estimators require a preliminary transformation of the component-wise maxima to new exponentially distributed r.v.s., i.e. $\xi_j = -\log \widehat{G}_1(X_j^*)$ and $\eta_j = -\log \widehat{G}_2(Y_j^*)$, where we use $\overline{\xi} = \frac{1}{m} \sum_{j=1}^{m} \xi_j$, $\overline{\eta} = \frac{1}{m} \sum_{j=1}^{m} \eta_j$. To transform the data, one has to estimate first the marginal DFs G_1 and G_2 by the component-wise maxima X_j^* and Y_j^*, e.g., by empirical DFs or parametric models. As such models one can take the Generalized Pareto distribution with the DF

$$\Psi_{\sigma,\gamma}(x) = \begin{cases} 1 - (1 + \gamma x/\sigma)^{-1/\gamma}, & \gamma \neq 0, \\ 1 - \exp(-x/\sigma), & \gamma = 0, \end{cases}$$

where $\sigma > 0$ and $x \geq 0$ if $\gamma \geq 0$ holds and $0 \leq x \leq -\sigma/\gamma$ if $\gamma < 0$ holds, or the Generalized Extreme Value distribution

$$H_\gamma(x) = \begin{cases} \exp\left(-\left(1 + \gamma \left(\frac{x-\mu}{\sigma} \right) \right)^{-1/\gamma} \right), & \gamma \neq 0 \\ \exp(-e^{-(x-\mu/\sigma)}), & \gamma = 0, \end{cases} \tag{2.8}$$

with $1 + \gamma (x - \mu)/\sigma > 0$.

The amount of these maxima may be moderate which may reflect on the accuracy of the DF estimates. The component-wise maxima may not be jointly observable, i.e. there are such pairs of maxima which do not exist in the original sample (the pairs are artificial). Then the question arises whether they should be excluded or not. Some observations may help to provide an answer:

- Artificial pairs do not affect the asymptotical distribution of the bivariate maxima and its Pickands' representation.
- Artificial pairs like those constructed from TCP-flow sizes and durations in a mobile network considered in [19] hug the vertical axis, see Figure 2.3, and hence, their components are independent (cf. [10]). Thus, they do not contribute to the dependence and can be excluded.
- Components of artificial pairs are not artificial itself and contribute to the margins.
- Excluding artificial pairs may lead to the reduction of the number of component-wise maxima and influences the trade-off between the bias and variance of the estimation.
- The accuracy of the estimation is more sensitive to the number of blocks rather than to the artificial pairs.

To check an estimate $\widehat{A}_m$ and to select the number of blocks m for a given sample size n, it is proposed in [20] to use a PP-plot $(\widehat{F_\chi}(\chi_{(i)}), i/m)$, $i = 1, \ldots, m$. Here

$$\widehat{F_\chi}(z) = \frac{z\left(1 + z - \frac{\widehat{A}'_m(\frac{1}{1+z})}{\widehat{A}_m(\frac{1}{1+z})}\right)}{(1+z)^2}$$

is the distribution of $\widehat{\chi} = \widehat{\xi}/\widehat{\eta}$, $\widehat{\xi} = -\log \widehat{G_1}(X^*)$ and $\widehat{\eta} = -\log \widehat{G_2}(Y^*)$. If the estimators of the A-function are not convex, they may be improved by taking a convex hull.

To test the required independence, the measure $2\,(1 - A(1/2))$ proposed in [30] can be used. For the estimator $\widehat{A}^C_m(t)$ it has the following approximate form

$$T_n \approx -\sqrt{n/0.342}\log \widehat{A}^C_m(1/2)$$

(see [3]). The null hypothesis on the independence of two r.v.s. is rejected at level α if T_n exceeds the quantile of order $1 - \alpha$ of the standard normal distribution. The problem of this test is determined by its slow convergence. In other words, its accuracy improves very slowly as the sample size increases. Hence, it can be unreliable for moderate sample size. Applying this test to the TCP-flow data of the mobile network considered in [20], for example, the null hypothesis regarding the independence of R and S, R and D has to be accepted with probability 99%.

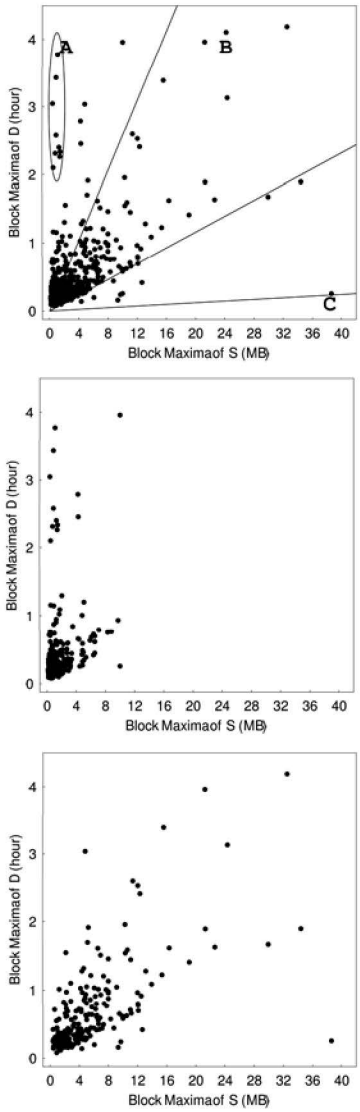

Figure 2.3 Top: a scatter plot of pairs of block maxima $(S_{(n),j}, D_{(n),j})$, $j = 1, \ldots, m$ when the block size is given by $n = 1\,000$ and $m = 610$. The lines, from bottom to top, indicate the access rates 384 kb/s (EDGE), 40.2 kb/s (a typical GPRS rate) and 9.05 kb/s (a minimum GPRS rate), respectively. Middle and bottom: artificial and true data points resulted from the maximization procedure, respectively.

2.5 Conclusions

Considering the statistical characterization of Internet traffic, three items have to be investigated:

(1) the preliminary detection of heavy tails;
(2) the dependence structure of the data;
(3) traffic modeling, model selection, estimation and evaluation.

For heavy-tailed models and, in particular, models with infinite variance, the classical statistical methods are not adequate or applicable and flexible enough. An example is provided by the sample autocorrelation function that has a specific form for heavy-tailed distributions with infinite variance and requires special confidence bounds. Moreover, we should further distinguish between methods that are valid for independent and dependent data.

In this paper we have presented simple methods to detect the heaviness of tails and dependence in data arising from Internet traffic. The exploratory techniques introduced in Section 2.2 like "the ratio of the maximum to the sum" or the group estimator of the tail index started from an i.i.d. assumption on the underlying data. Their interpretation may become hazardous when they are applied to the non-iid case.

In Section 2.3 we have discussed some methods to test the dependence in univariate data and focussed on the estimation of the extremal index. In Section 2.4 techniques to determine the dependence in bivariate data have been stated. They have been illustrated by the application of Pickands' function to TCP-flow data. The estimation of Pickands' function assumes that the underlying (size, duration) pairs of flows are independent, too.

In conclusion, we are convinced that the presented statistical techniques and examples illustrating their application to real Internet data provide a useful guideline for the rigorous teletraffic analysis of various features arising from the measurements of Internet traffic at the time scales of the session and flow levels.

Acknowledgement

The research was partly supported by the FP6-NoE-project EuroFGI under contract 028022.

References

[1] J. Beirlant, Y. Goegebeur, J. Teugels and J. Segers, *Statistics of Extremes: Theory and Applications*, Wiley, Chichester, West Sussex, 2004.

[2] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd edition, Springer, New York, 1991.

[3] P. Capéraà, A.-L. Fougères and C. Genest, A nonparametric estimation procedure for bivariate extreme value copulas, *Biometrika*, vol. 84, pp. 567–577, 1997.

[4] M. E. Crovella and A. Bestavros, *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes*, ACM Sigmetrics, pp. 226–244, 1996.

[5] Y. Davydov, V. Paulauskas and A. Račkauskas, More on P-stable convex sets in Banach spaces, *J. Theoret. Probab.*, vol. 13, no. 1, pp. 39–64, 2000.

[6] P. Embrechts, C. Klüppelberg and T. Mikosch, *Modeling Extremal Events*, Springer, Berlin, 1997.

[7] Eurescom Project P1112, New dimensions – Network dimensioning based on modeling of Internet traffic. Traffic characteristics and statistical estimation, Technical Report D8, Heidelberg, June 2003.

[8] C. A. T. Ferro and J. Segers, Inference for clusters of extreme values, *J. Royal Stat. Soc., Ser. B*, vol. 65, pp. 545–556, 2003.

[9] P. Hall and N. Tajvidi, Distribution and dependence-function estimation for bivariate extreme-value distributions, *Bernoulli*, vol. 6, pp. 835–844, 2000.

[10] F. Hernandez-Campos, J. S. Marron, S. I. Resnick, C. Park and K. Jeffay, Extremal dependence: Internet traffic applications, *Stochastic Models*, vol. 21, no. 1, pp. 1–35, 2005.

[11] B. M. Hill, A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, vol. 3, pp. 1163–1174, 1975.

[12] J. Jurečková and J. Picek, A class of tests on the tail index, *Extremes*, vol. 4, pp. 165–183, 2001.

[13] H. Kettani and J. A. Gubner, A novel approach to the estimation of the hurst parameter in self-similar traffic, in *Proceedings of IEEE Conference on Local Computer Networks*, Tampa, Florida, 2002.

[14] U. R. Krieger, N. M. Markovitch and N. Vicari, Analysis of World Wide Web traffic by nonparametric estimation techniques, in *Performance and QoS of Next Generatione Networking*, K. Guto et al. (Eds.), Springer, London, pp. 67–83, 2001.

[15] U. R. Krieger (Ed.), *New Mathematical Methods, Algorithms and Tools for Measurement, IP Traffic Characterization and Classification*, IST-FP6 NoE EuroFGI, Contract No. 028022, Deliverable D.WP.JRA.5.1.1, December 2007.

[16] U. R. Krieger (Ed.), *Achievements on Measurements, IP Traffic Characterization, Classification and Statistical Methods*, IST-FP6 NoE EuroFGI, Contract No. 028022, Deliverable D-WP-JRA-5.1-2, March 2008.

[17] N. M. Markovich, High quantile estimation for heavy-tailed distributions, *Performance Evaluation*, vol. 62, nos. 1–4, pp. 178–192, 2005.

[18] N. M. Markovich, *Nonparametric Estimation of Univariate Heavy-Tailed Data. Research and Practice*, J. Wiley & Sons, Chichester, 2007.

[19] N. M. Markovich and J. Kilpi, Bivariate statistical analysis of TCP-flow sizes and durations, in *Proceedings Stochastic Performance Models for Resource Allocation*

in Communication Systems, Amsterdam, 8–10 November 2006, pp. 47–50, 2006, http://www.cwi.nl/events/2006/StoPeRa/.

[20] N. Markovich and J. Kilpi, Bivariate statistical analysis of TCP-flow sizes and durations, *Annals of Operations Research*, to be published, 2008.

[21] N. M. Markovitch and U. R. Krieger, The estimation of heavy-tailed probability density functions, their mixtures and quantiles, *Computer Networks*, vol. 40, no. 3, pp. 459–474, 2002.

[22] N. M. Markovich, A. Undheim and P. Emstad, Classification of slice-based VBR video traffic and estimation of link loss by exceedance, *Computer Network*, 2009.

[23] R. van de Meent and M. Mandjes, Evaluation of 'user-oriented' and 'black-box' traffic models for link provisioning, in *Proceedings of 1st Conference on Next Generation Internet Design and Engineering*, Rome, Italy, 2005, IEEE, Piscataway, NY, pp. 380–387, 2005.

[24] M. Nabe, M. Murata and H. Miyahara, Analysis and modeling of World Wide Web traffic for capacity dimensioning of Internet access lines, *Performance Evaluation*, vol. 34, pp. 249–271, 1998.

[25] C. Neves and I. Fraga Alves, The ratio of maximum to the sum for testing super heavy tails, in *Advances in Mathematical and Statistical Modeling*, B. C. Arnold, N. Balakrishnan, J. M. Sarabia and R. Minguez (Eds.), Birkhäuser, Boston, pp. 181–194, 2008.

[26] A. Pacheco, C. Pascoal and M. R. de Oliveira, Analysis of internet traffic flows using the truncated bivariate normal distribution, Technical Report, 2007, see [15].

[27] S. I. Resnick, *Heavy-Tail Phenomena. Probabilistic and Statistical Modeling*, Springer, New York, 2006.

[28] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York, 1986.

[29] R. L. Smith and I. Weissman, Estimating the extremal index, *J. Royal Stat. Soc., Ser. B*, vol. 56, no. 3, pp. 515–528, 1994.

[30] J. A. Tawn, Bivariate extreme value theory: Models and estimation, *Biometrika*, vol. 75, no. 3, pp. 397–415, 1988.