Statistical Analysis and Modeling of Peer-to-Peer Multimedia Traffic

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

[1] Institute of Control Sciences, Russian Academy of Sciences
Profsoyuznaya Str. 65, Moscow 117997, Russia
markovic@ipu.rssi.ru
[2] Faculty Information Systems and Applied Computer Science
Otto-Friedrich-Universität, D-96052 Bamberg, Germany
udo.krieger@ieee.org

Abstract. We study peer-to-peer packet traffic arising from passive VoIP and video measurements that are generated by Skype and IPTV clients. We provide a common methodology for the statistical characterization of the packet flows, discuss the user's satisfaction and load estimation. Two main ideas are used in our analysis. Due to the dependence of the data we first partition the observations into independent blocks and deal further with these block-wise independent data. Secondly, loss is generated by packet lengths which exceed the channel capacity in a time unit if the inter-arrival times coincide with this time unit. If the inter-arrival times are random, loss is generated by the lengths of those packets corresponding to transmission rates that exceed the channel capacity. Our methodology is demonstrated by individual Skype flows and the aggregated flow of video packets exchanged with a mobile peer of a SopCast session.

Keywords: Peer-to-peer traffic characterization, multimedia packet traffic, Skype, IPTV, SopCast.

1 Introduction

In recent years, peer-to-peer (P2P) multimedia applications like Skype, IPTV and on-line games have become a powerful service platform for the generation and transport of voice and video over IP. Due to its free access Skype, for instance, is now a real competitor of the traditional telephony services and has millions of customers.

Due to randomly appearing peers the main feature of a P2P application is determined by the random structure of its overlay network. The peers are both receivers and senders of chunks of information at the same time. They determine random transmission processes of this information to a cloud of receivers. The management and control of the traffic dynamics is complicated since it depends on this randomness in the overlay network and the transmission processes. To improve the understanding of the transport mechanisms in P2P networks, we

investigate the statistical properties of P2P multimedia traffic at the interaction level and time scale of the packet layer.

Regarding Skype traffic different types of media like voice, video, and text messages are transferred by a client. Considering IPTV life sessions, video and voice streams are transmitted by the P2P network. Currently, there are several important P2P TV applications, including PPlive, PPStream, SopCast and TVAnts. Many authors have already tried to classify the VoIP traffic of Skype sessions and the packet flows of IPTV sessions regarding the used applications, applied encoding schemes etc., see for example [6], [12], [22]. These studies concern both the packet and flow level characterization of the monitored traffic streams.

In our study we do not intend to classify the gathered P2P traffic. But we study passive measurements of VoIP flows arising from a Skype client and video traffic generated by an IPTV session at the Ethernet packet layer. We focus on the inter-arrival time (IAT) and packet length (PL) processes and provide a common methodology for the statistical characterization, the analysis of the user's satisfaction and the load estimation of the packet transmission. This approach is possible since the analysis of the packet transmission over IP has a common foundation irrespective of the P2P video or voice transfer.

The characterization implies that we have to investigate whether the traffic is stationary, long- or short-range dependent or independent, self-similar (i.e. scale invariant), and heavy-tail distributed. The analysis can be done by a common and rigorous mathematical methodology. It is illustrated by examples of an aggregated IPTV traffic flow to a mobile peer and the VoIP traffic in a WLAN environment between two Skype users.

Since the data are mostly dependent, we have to partition it into independent blocks and to deal with representatives of these blocks like maxima, minima and averages just like with independent data. Particularly, this procedure allows us to fit the distribution of the maximum of the IATs between packets.

Another important question of our study concerns the user's satisfaction. It is determined at the packet layer by the loss and delay of transmitted packets. Both impact on the quality of service (QoS) and the user's quality of experience (QoE). The extremes (i.e. maximal and minimal values) of the IATs between delivered packets and of the PLs influence on the speech and image perception more than the non-extremal values of these indices. Hence, we model the distribution of the maximal IAT between packets and find its quantiles as indices of the quality. Besides, we propose the mean byte loss, the mean delivery time variation of packets per cluster and the quantiles of lossless periods as new indices of the quality.

The rest of the paper is organized as follows. In Section 2 the used data sets of the P2P multimedia flows are described. In Section 3 the common methodology to detect stationarity, long-range dependence (LRD), self-similarity and the heaviness of tails is presented. In Section 4 the partitioning of the observations into independent blocks as preliminary tool for the analysis of dependent data is explained. In Section 5 we present indices of the user's satisfaction during

a packet transmission and illustrate them by means of Skype traffic data. In Section 6 the estimation of the offered traffic load in a finite time interval is presented. In Section 7 the Extreme Value Distribution and its quantiles describing the maximum of the IATs between transmitted packets are evaluated for IPTV data. Finally, some conclusions are presented.

2 Description of the Multimedia Packet Traffic

Actually, we use in our study the inter-arrival times (IATs) between packets and the packet lengths (PLs) as the main source of information. In the following the sequence of n IATs between the packets of a multimedia traffic flow is denoted by $X_1, X_2..., X_n$ and $Y_1, Y_2..., Y_n$ are the associated PLs. Our proposed methodology is demonstrated by means of two illustrative data sets containing peer-to-peer video and voice-over-IP (VoIP) traffic.

2.1 Description of the VoIP Data

To reveal the features of our statistical techniques for multimedia traffic characterization, we have used P2P VoIP traffic generated by Skype clients, cf. [26]. Due to the peer-to-peer character of Skype and the random nature of its overlay network relaying the generated packet flows, it is in general difficult to monitor the traffic between two communicating hosts along a path in the overlay network. Hence, one must gather the Skype packet traffic related to a particular site, cf. [8].

We have also followed this approach. Communication sessions between two Skype clients within a LAN test bed and their encoded voice samples have been gathered by means of Wireshark at Otto-Friedrich University Bamberg in 2006 (see Fig. 1 taken from [15], [16, Fig. 1]). The collected PLs and IATs between Ethernet packets of a representative single Skype flow will be used as first illustrative data set in our study. The latter flow has been generated based on a mixture of short representative sessions of monologues, dialogs and music clips with German and English male and female speakers and lasts 135 seconds. The resulting unidirectional VoIP packet stream has been isolated in a pre-processing phase. The variable bit rate wideband Internet Speech Audio Codec (iSAC) has been applied by the clients as basic voice encoding scheme with a sampling frequency of 16 kHz. It is able to respond to varying network conditions and generates variable data rates ranging from 10 to 32 kbps.

This data set illustrates the typical features of Skype flows in current home environments. It has been arising from a transmission path with several wired and two wireless links. Here both the sending mobile client (mobile host 2 at 192.168.182.22) and the receiving mobile client (mobile host 1 at 192.168.1.4) are first traversing private IEEE802.11 WLAN segments 1 and 5, respectively, with DSL attachments to the public Internet and then an Internet path 2,3,4 including

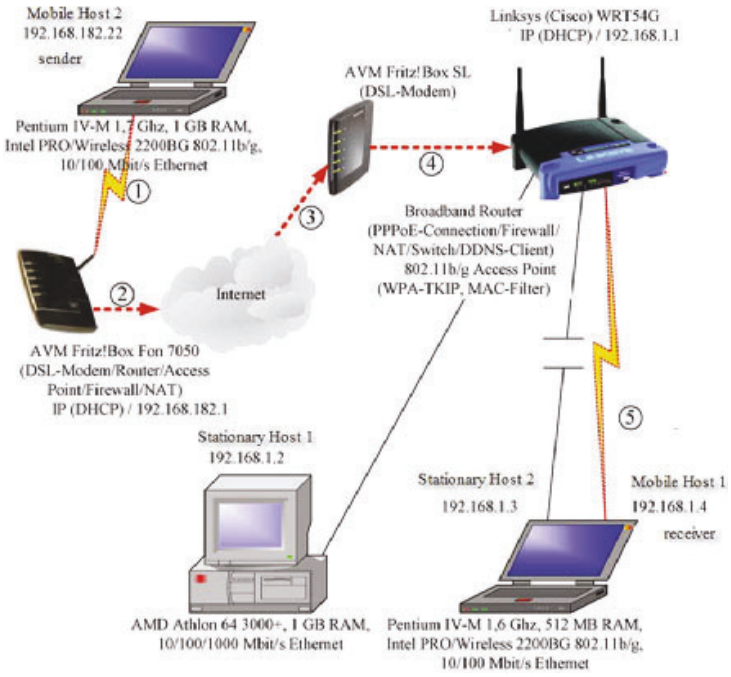


Fig. 1. LAN test bed for voice over IP communication by Skype clients (see [16, Fig. 1])

a tier-1 carrier exchange point between Telefonica's and Deutsche Telekom's ISP networks (see Fig. 1, cf. [16, Fig. 1]). Therefore, this network path can be considered as typical VoIP over WLAN environment that a majority of Skype users traverse today between two private homes.

To evaluate the load and delivery variation profile, we also need the extremes of the PLs and IATs within independent subsets (called blocks) of data, see Section 4. All descriptive statistics of these random variables (r.v.s) arising from our representative VoIP data set are stated in Table 1.

2.2 Description of the IPTV Data

The second illustrative data set contains P2PTV traces generated by the P2P IPTV system SopCast [27]. A comprehensive measurement study of a typical IPTV home scenario including a wireless access to the Internet has been performed during the second quarter of 2009 by the Computer Networks Laboratory of Otto-Friedrich University Bamberg, Germany.

In this wireless scenario the SopCast client is running on a desktop IBM Thinkcentre with 2.8 GHz Intel Pentium 4 processor, 512 MB RAM, and Windows XP Home. It is attached by a Netgear WG111 NIC operating the IEEE802.11g MAC protocol over a wireless link to the corresponding ADSL router acting as gateway to the Internet.

Watching a popular sport channel, representative traces arising from sessions of 30 minutes have been gathered by Wireshark at the mobile host. The descriptive statistics of a representative aggregated flow to the observed SopCast client are stated in Tab. 2.

Table 1. Description of the VoIP data arising from a Skype packet flow

R.V.	Sample Size	Min	Max	Mean	StDev	Skewness	Kurtosis
Inter-arrival times (sec)	4605	$1.9 \cdot 10^{-5}$	$2.01 \cdot 10^{-1}$	$3.1 \cdot 10^{-2}$	$8.635 \cdot 10^{-3}$	5.183	79.75
Packet lengths (bytes)	4605	45	284	160.27	25.808	−0.921	2.32
Maxima of inter-arrival times (sec)	72	$5.8 \cdot 10^{-2}$	$2.01 \cdot 10^{-1}$	$7.3 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	3.622	14.253
Minima of inter-arrival times (sec)	72	$1.9 \cdot 10^{-5}$	$9.1 \cdot 10^{-2}$	$3.3 \cdot 10^{-2}$	$6.816 \cdot 10^{-4}$	0.028	−1.56
Maxima of packet lengths (bytes)	72	85	284	197.69	1688	−1.405	6.857

Table 2. Description of the IATs between packets in seconds and the block maxima corresponding to IAT blocks of size 400 (IAT400) arising from the aggregated flow to the observed peer

R.V.	Sample Size	Min	Max	Median	Mean	StDev	Skewness	Kurtosis
IAT	$6.553 \cdot 10^4$	$2.1 \cdot 10^{-5}$	0.625	$5.58 \cdot 10^{-4}$	$4.934 \cdot 10^{-3}$	0.016	13.56	313.087
IAT400	163	0.021	0.625	0.105	0.138	0.102	1.773	4.193

3 Statistical Characterization of P2P Packet Flows

3.1 Detection of Stationarity

Before applying any statistical analysis method to time series it is the first step to check whether the data are stationary. In practice it is not realistic to observe a pure stationary process. In this case we can partition the observations at our disposal into homogeneous sequences of data which are approximately stationary.

The weak stationarity of a stochastic process $\{X_t, t \geq 0\}$ requires that the first two moments and the autocorrelation function (ACF) do not change in time, i.e. $\mu = \mathbb{E}(X_t)$, $\sigma^2 = \mathrm{Var}(X_t)$, and the ACF between X_t and X_s only depends on the difference $|t - s|$. Non-stationarity is equivalent to the presence of a deterministic or stochastic trend in the data.

All tests on stationarity, e.g., Cochran's test [9], the runs test [3], the R/S method [23] (see also Section 3.3), are based on a partitioning of the data into independent blocks and the comparison of averages, deviations from averages, standard deviations or other statistical characteristics calculated by means of these blocks.

The runs test, for instance, recommends to divide the time series into equal-sized time intervals, to compute a mean value for each interval and to count the number of runs of the mean values above and below the median value of the series. Then one has to compare the calculated number of counts with the value that one would expect if the observations were independent of each other.

In [5] the R/S test is applied to detect the presence of deterministic trends in the data. A survey of recent methods is provided by [10]. However, all these tests have own constraints and drawbacks. An important constraint is determined by the independence of data blocks that is difficult to achieve. Therefore, we consider here some rough tools to check whether the mean and variance do not change in time.

Let $\{X_i, i = 1, 2, ..., n\}$ denote the original time series. Regarding the partitioned data we then calculate the averages and variances within each block of size m numbered by the integer $k = 1, 2, ..., [n/m]$

$$X^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} X_i,$$

$$V^{(m)}(k) = \frac{1}{m-1} \sum_{i=(k-1)m+1}^{km} \left(X_i - X^{(m)}(k)\right)^2$$

and the sample variance of $X^{(m)}(k)$

$$\widehat{\mathrm{Var}} X^{(m)} = \left[\frac{m}{n}\right] \sum_{k=1}^{[n/m]} \left(X^{(m)}(k)\right)^2 - \left(\left[\frac{m}{n}\right] \sum_{k=1}^{[n/m]} X^{(m)}(k)\right)^2. \qquad (1)$$

To test the stationarity with regard to the homogeneity of the mean, we check the difference of the sample variances

$$D(m) = \widehat{\mathrm{Var}} X^{(m)} - \widehat{\mathrm{Var}} X^{(m-1)}$$

for successive values of m, cf. [23].

To check the homogeneity of the variances, one can apply Cochran's test. The idea of this test is simply to calculate the ratios

$$G_k = V^{(m)}(k) / \sum_{i=1}^{[n/m]} V^{(m)}(i), \qquad k = 1, 2, ..., [n/m].$$

Then it is the objective to select the maximal value $G_{max} = \max_k G_k$ among them and to compare it with the quantiles of the distribution of $G_{\max}$ for a required level. Choran's test requires the independence and normality of the block data. In Section 4 we discuss a method to partition our data into independent blocks. However, the normality of the data over the blocks is not always fulfilled. According to the Central Limit Theorem the average over the blocks can be asymptotically normal distributed if the representatives of the blocks are independent and their second moment is finite. In Section 3.2 we discuss a way to detect the presence of heavy tails in the data and to understand how many moments of the distribution are finite. Note, that the normal distribution is light-tailed.

Moreover, it is known that it is difficult to distinguish between stationary processes with a long memory and non-stationary processes (cf. [4, Chap. 7.4, p. 141f]). We show in Section 3.3 that both illustrative data sets exhibit a long-range dependent (LRD) behavior.

Example 1: We check first the stationarity of the Skype packet data, namely, the IATs and PLs (regarding their description see Section 2.1).

The means of the IATs and PLs do not change much (see Fig. 2(b), 2(d), cf. also [15]). One cannot conclude definitely from the visual analysis that the variances do not change much in order to expect the non-stationarity of the IATs and PLs (see Fig. 2(a), 2(b), 2(e), 2(f), cf. also [15]).

Applying Cochran's test, we get $G_{\max} = 0.19$ regarding the IATs and $G_{\max} = 0.114$ for the PLs, cf. [15]. For $m = 145$ and $[n/m] = 31$ the 5% quantile of $G_{\max}$ is equal to 0.0457. The null hypothesis regarding the equality of the variances should be rejected since the calculated values $G_{\max}$ of the IATs and PLs are larger than the bound 0.0457. Nevertheless, this conclusion may be unreliable due to the deviation of the data from normality. The latter assumption constitutes the main constraint of this test. Due to a positive kurtosis the IATs and PLs are not normal distributed. Their non-zero skewness' indicate that the distributions of both characteristics are asymmetric, see Tab.1.

To proceed in a mathematically rigorous way, we partition our data into independent blocks and check the stationarity of their representatives by an inversion test, cf. [3]. In Section 4.2 we consider the techniques to select such blocks and apply it to the Skype data of the example.

The null hypothesis states that the underlying sequence contains independent stationary random observations, i.e., a trend does not exist. For this purpose one calculates the statistic

$$A = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}(X_i > X_j).$$

The hypothesis is accepted at level $\alpha = 0.05$ if $A_{n,1-\alpha/2} < A \leq A_{n,\alpha/2}$ holds, where $A_{n,1-\alpha/2}$ and $A_{n,\alpha/2}$ are quantiles of the distribution function (DF) of A. The bounds of A are given by $[1014, 1400]$.

We investigate the stationarity and independence of the maxima of the PLs and the increments $X_t - X_{t-1}$ of the maxima and minima of the IATs between Skype packets in the blocks. Since the values of A fall into the mentioned interval (see Tab. 3), the null hypothesis should be accepted for the maxima of PLs as well as the maxima and minima of the IATs.

3.2 Detecting the Heaviness of Tails of the Distributions

Let $F(x)$ denote the distribution function (DF) of the underlying r.v. X, e.g., the IAT. Roughly speaking, heavy-tailed distributions are those long-tailed distributions whose tails decay to zero slower than an exponential tail. The tail is determined by the function $1 - F(x)$. Regularly varying distributions with

$$1 - F(x) = x^{-\alpha}\ell(x), \tag{2}$$

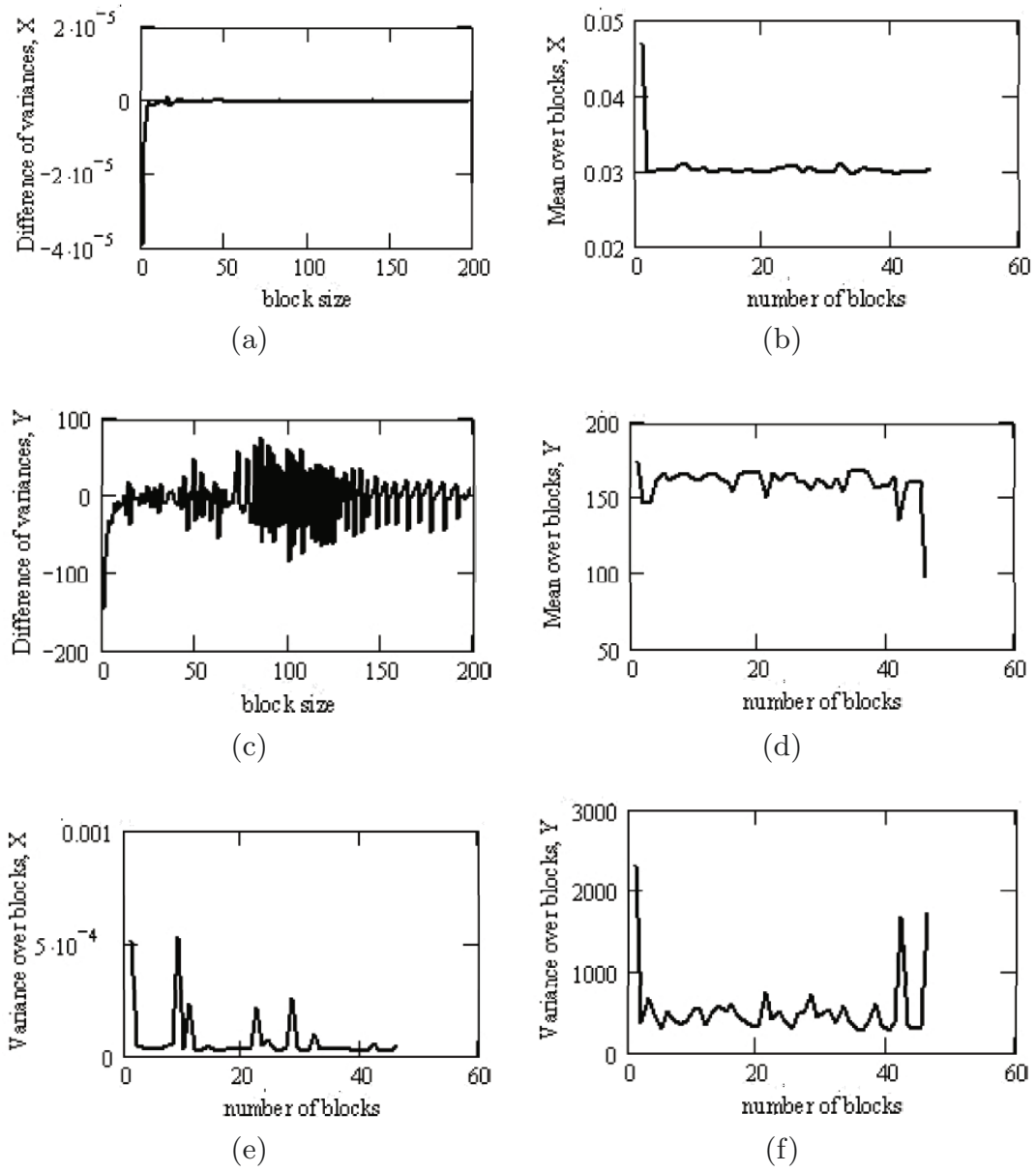


Fig. 2. The differences of variances $D(m)$ and the means $X^{(m)}(k)$ within each block of the inter-arrival times between Skype packets (a) and (b) and the lengths of Skype packets (c) and (d); the size of blocks $m = 100$ was chosen for (b) and (d); the variance $V^{(m)}(k)$ within each block of inter-arrival times (e) and packet lengths (f)

where $\ell(x)$ is a slowly varying function with the property $\lim_{x\to\infty} \ell(xt)/\ell(x) = 1$ for any $t > 0$, constitute the widest class of heavy-tailed distributions.

The tail index α or its reciprocal $\gamma = 1/\alpha$ called the extreme value index (EVI) show the shape of the tail of the distribution F. A positive sign of α implies that the distribution is heavy-tailed. The smaller α the heavier is the tail. Further, α indicates the number of finite moments, namely, $\mathbb{E}x^\beta < \infty$ holds for $\beta < \alpha$ if the distribution is regularly varying. In contrast to light-tailed distributions, not all moments of heavy-tailed ones are finite. All moments are infinite for super-heavy-tailed distributions.

Table 3. Inversion test results (cf. [16, Table V])

	Maxima of Packet Length	**Increments of Inter-arrival Times**	
		Maxima	*Minima*
A	1358	1294	1249

There are several rough tools that allow us to distinguish between light and heavy tails, see, e.g., [14] for a survey. Here we describe the most evident methods, namely, the estimation of the tail index and the mean excess function.

To estimate the EVI γ, we use the popular Hill's estimator

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

Here $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ denote the order statistics of the sample $\{X_i, i = 1,\ldots,n\}$ and k is a smoothing parameter. One can select k corresponding to the stability interval of the Hill's plot $\{(k, \widehat{\gamma}^H(n,k)), k = 1,\ldots,n-1\}$.

One can estimate γ by a bootstrap procedure as an automatic method. This scheme implies the averaging of the Hill's estimates constructed over bootstrap re-samples that are taken from the underlying sample with repetitions, see [14, pp. 22–25].

There are numerous estimators of the tail index, but many of them are very sensitive to dependence in the data. The Hill's estimator, however, can be applied to dependent data subject to specific mixing conditions, cf. [19].

The mean excess function

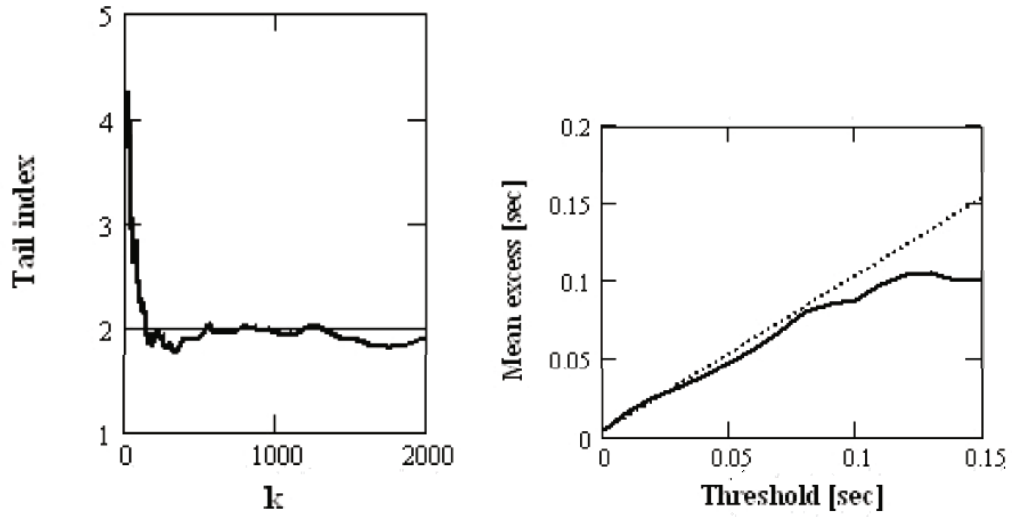
$$e(u) = \mathbb{E}(X - u | X > u) \tag{3}$$

provides another method that can indicate a heavy tail. The increase (or decrease) of $e(u)$ implies a heavy-tailed (or light-tailed) distribution. Its constant value corresponds to an exponential distribution. Its linear increase indicates a Pareto-like distribution. Usually, the sample mean excess function

$$e_n(u) = \sum_{i=1}^{n}(X_i - u)\mathbb{1}(X_i > u) / \sum_{i=1}^{n}\mathbb{1}(X_i > u),$$

is calculated, where $\mathbb{1}(A)$ is the indicator function of the event A. Relatively large values of u are usually not considered due to the few observations exceeding these thresholds. It leads to unreliable estimates $e_n(u)$.

Example 2: Let us consider the IPTV IAT data as a typical example. Regarding the stability interval of the Hill's plot, one can find the corresponding value of the tail index $\hat{\alpha} \approx 2$, see Fig. 3(a). The bootstrap method over 200 bootstrap re-samples taken from the IPTV IAT sample with repetitions yields a similar value $\hat{\alpha} = 1.883$, cf. [18]. Assuming a regular varying DF, this value implies that only the first moment of the IAT distribution is finite and, hence, the distribution is heavy-tailed.



(a) (b)

Fig. 3. Estimation of the tail index of the IPTV IATs by the reciprocal of Hill's estimator against the number of the largest order statistics k (a) and the mean excess function of the IPTV IATs against the threshold u (b)

Since the Hill's estimate may be corrupted by dependence if necessary mixing conditions are not fulfilled, we also estimate the bootstrapped Hill's estimate of the independent block maxima of the IPTV IATs calculated over 500 bootstrap re-samples. $m = 400$ has been selected as block size to provide independent blocks, see [18]. Then we have obtained $\hat{\alpha} = 3.39$. This implies that the first three moments of this distribution are finite. The distribution of the IAT block maxima is heavy-tailed.

By Fig. 3(b) one can conclude that the IPTV IAT distribution is a mixture of a Pareto-like and an exponential distributions since $e_n(u)$ increases almost linear up to the threshold $u = 0.12$ and is almost constant beyond 0.12. The latter is the 99.8% empirical quantile of the IATs.

3.3 Detection of Long-Range Dependence

Long-range dependence (LRD) of a time series $\{X_t, t = 1, \ldots, n\}$ means that the ACF $\rho_X(h) = \mathbb{E}\left((X_t - \mu)(X_{t+h} - \mu)\right)/\sigma^2$, $\mu = \mathbb{E}(X_t), \sigma^2 = \mathrm{Var}(X_t)$, remains sufficiently large in magnitude over a long period of time. The values of the ACF can be small, but in case of LRD their cumulative effect is significant, i.e., $\sum_{h=0}^{\infty} |\rho_X(h)| = \infty$.

The dependence structure may be derived by calculating the sample ACF and the extremal index and executing Portmanteau tests. An LRD property may be detected by an estimation of the Hurst parameter.

Estimation of the Autocorrelation Function. The classical sample ACF is calculated by the formula

$$\widehat{\rho}(h) = \frac{\sum_{t=1}^{n-h} \left(X_t - \overline{X}_n\right)\left(X_{t+h} - \overline{X}_n\right)}{\sum_{t=1}^{n} \left(X_t - \overline{X}_n\right)^2}$$

with $\overline{X}_n = 1/n \sum_{i=1}^{n} X_i$. For normal distributed characteristics one may conclude that independence or short-range dependence occurs if the ACF is dying after a few lags h inside the 95% Gaussian confidence window $\pm 1.96/\sqrt{n}$. For non-Gaussian r.v.s this conclusion may be wrong.

Note that the ACF does not exist when the variance is infinite. In this case one can use the modified estimate without a centering by the sample mean,

$$\widehat{\rho}_n(h) = \sum_{t=1}^{n-h} X_t X_{t+h} / \sum_{t=1}^{n} X_t^2,$$

instead of $\rho_X(h)$, cf. [20]. This estimate may be unreliable for non-linear processes.

In practice one can use the classical sample ACF even if the distribution is heavy-tailed, cf. [20, p. 349]. However, it is difficult to check the null hypothesis that the data stems from an independent sample because the confidence intervals of the ACF cannot be defined easily for heavy-tailed distributions in contrast to the Gaussian bounds of the $\widehat{\rho}(h)$, see [7]. The conclusion is that one has to apply additional tests to check the dependence.

Example 3: We consider the sample ACFs of our illustrative Skype and IPTV data sets. The ACFs of both the IATs and PLs of a Skype flow are small but do not decrease at large lags, see Fig. 4(a), 4(b) (cf. also [15]). This property implies that both the IATs and PLs of a Skype packet flow may be LRD.

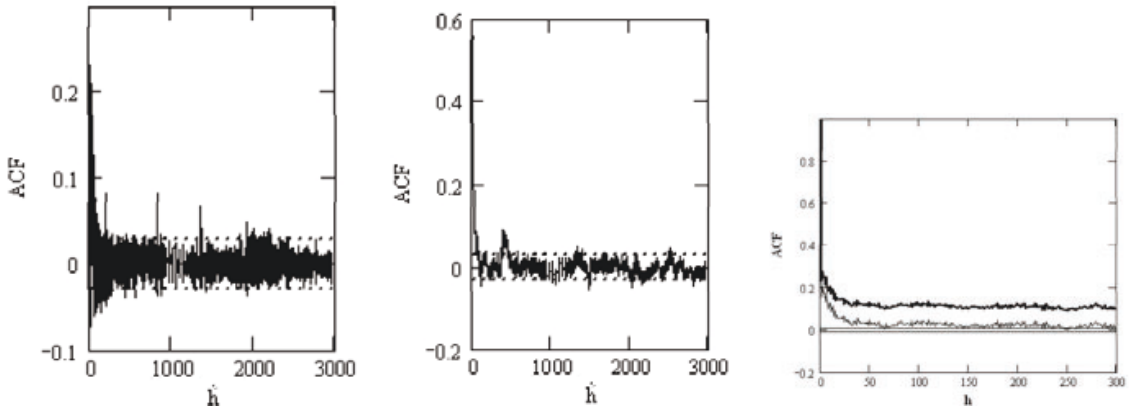


Fig. 4. The sample ACF $\widehat{\rho}(h)$ of the IATs between Skype packets (a) and of the lengths of Skype packets (b). The sample ACFs of the IATs between IPTV packets $\widehat{\rho}(h)$ (thin line) and $\widehat{\rho}_n(h)$ (solid line) (c). All estimates are shown with Gaussian 95% confidence intervals with the bounds $\pm 1.96/\sqrt{n}$.

Since the variance of the IATs of IPTV video data is infinite (see example 2 in Section 3.2) we consider both sample ACFs $\widehat{\rho}(h)$ and $\widehat{\rho}_n(h)$. They do not decrease as the lag h increases and do not remain within the Gaussian 95% confidence interval, see Fig.4(c) (cf. also [18]). Both facts may confirm the LRD of the IPTV IATs.

Portmanteau Tests. We consider two Portmanteau tests, namely, the Ljung-Box test and Runde's test. The first one is appropriate if the variance of the underlying r.v. is finite whereas the second is valid for time series with infinite variance. These tests check the null hypothesis regarding the independence of the underlying data.

According to the Ljung-Box test the test statistic

$$Q = n(n+2)\sum_{j=1}^{h} \hat{\rho}^2(j)/(n-j)$$

has approximately a chi-square distribution with h degrees of freedom, cf. [7], [13]. The i.i.d. hypothesis should be rejected at level η if $Q > \chi^2_\eta(h)$ holds, where $\chi^2_\eta(h)$ is the ηth quantile of the chi-square distribution with h degrees of freedom, i.e., $Pr\{\chi^2 > \chi^2_\eta(h)\} = \eta$, $0 < \eta < 1$.

Given the tail index $1 < \alpha < 2$, Runde's test [21] uses the test statistic

$$Q_R = (n/\ln n)^{2/\alpha} \sum_{j=1}^{h} \hat{\rho}^2(j).$$

The condition $1 < \alpha < 2$ implies that the second moment of the distribution is infinite, see Section 3.2. The quantiles of the limiting stable distribution of Q_R can be found in [21]. Some of them are given in Tab. 4, cf. [18]. To use Runde's test we have to estimate first the tail index (see Section 3.2 for corresponding methods). Since this test is valid for symmetric r.v.s, one has to construct a new symmetric r.v. based on the underlying r.v.s $\{X_i\}$. For this purpose we consider $Y_i = s_i X_i$, where s_i is a discrete r.v. that takes the values $+1$ and -1 with the probabilities 0.5. Y_i has the same tail index α as X_i since the DF of Y_i is determined by

$$\mathbb{P}\{Y_i \le x\} = 1/2\mathbb{P}\{|s_i X_i| \le x\} = 1/2\mathbb{P}\{X_i \le x\}.$$

Now we can check the independence of Y_i. If $\{Y_i\}$ are independent then $\{X_i\}$ are independent, too, since

$$\begin{aligned} 1/2^n\ & \mathbb{P}\{X_1 \le x_1\}...\mathbb{P}\{X_n \le x_n\} = \mathbb{P}\{Y_1 \le x\}\ldots\mathbb{P}\{Y_n \le x\} \\ = \ & \mathbb{P}\{Y_1 \le x_1, ..., Y_n \le x_n\} = \mathbb{P}\{s_1X_1 \le x_1, ..., s_nX_n \le x_n\} \\ = \ & 1/2^n\mathbb{P}\{|s_1X_1| \le x_1, ..., |s_nX_n| \le x_n\} = 1/2^n\mathbb{P}\{X_1 \le x_1, ..., X_n \le x_n\}, \end{aligned}$$

holds, cf. [16].

Table 4. Results of the Ljung-Box and Runde's test for IPTV data

Data	Lags	$\mathbf{Q_R}$	$\mathbf{Q_h(0.05)}$	Data	Lags	Q	$\chi^2_{0.05}(\mathbf{h})$
IAT	2	$4.01 \cdot 10^3$	13.53	IAT block	10	13.273	18.3
	3	$5.917 \cdot 10^3$	16.32	maxima	20	25.515	31.4
	4	$7.82 \cdot 10^3$	18.28	$\{X_i^{400}\}$	30	40.696	43.8
	5	$9.344 \cdot 10^3$	19.17				

Example 4: We apply Runde's test to the IATs between IPTV packets, see Section 2.2. The latter test is appropriate since the variance of the IAT distribution is infinite, see example 2 in Section 3.2. Since the values of Runde's statistic Q_R with the estimated $\hat{\alpha} = 1.883$ exceed the critical values $Q_h(0.05)$ of the limiting distribution of Q_R for the 0.05−level given in [21], the null hypothesis regarding the independence of IPTV IATs should be rejected, see Tab. 4.

In contrast to that, the block maxima $\{X_i^{400}\}$ of the IPTV IATs calculated over equal-sized blocks of size 400 may be independent. We can apply the Ljung-Box test to this data set since the estimate of its tail index is equal to $\hat{\alpha} = 3.39$ which is larger than 2, see example 2 in Section 3.2. Hence, the variance is finite. Since the values Q do not exceed $\chi^2_{0.05}(h)$, the null hypothesis regarding the independence of these block maxima should be accepted, see Tab. 4.

Estimation of the Extremal Index. To check the dependence additionally, a constant $\theta \in [0, 1]$ of the process known as its extremal index is calculated. θ shows the change in the limiting distribution of the sample maximum due to dependence in the process. According to the theory of extremes typically

$$\mathbb{P}\{M_n \leq u\} \approx \mathbb{P}^\theta\{\widetilde{M}_n \leq u\} = F^{n\theta}(u) \tag{4}$$

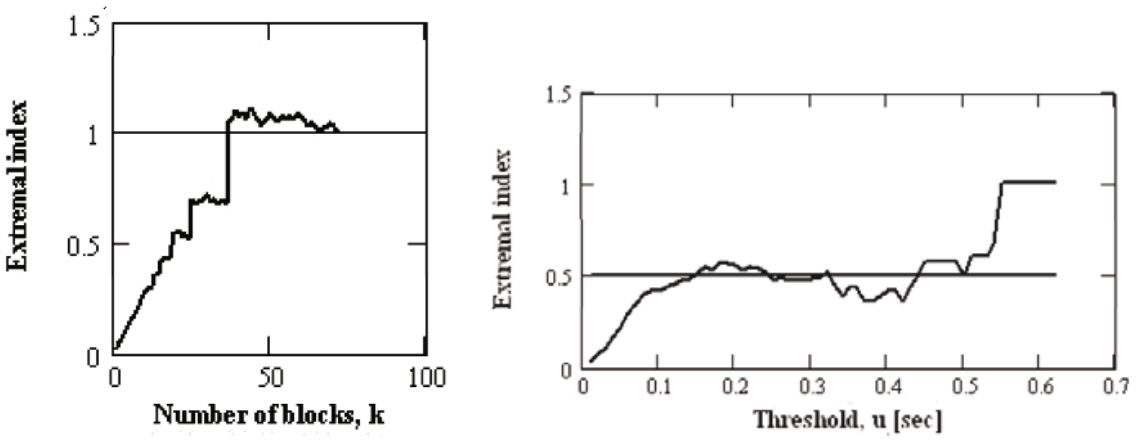
holds for sufficiently large n and u. Here M_n and $\widetilde{M}_n$ are the maximum of the sequence of dependent r.v.s $\{X_1, ..., X_n\}$ and the sequence of associated independent r.v.s $\{\widetilde{X_1}, ..., \widetilde{X_n}\}$ with the same DF $F(x)$, cf. [2]. For independent identically distributed (i.i.d.) sequences $\theta = 1$ holds.

The dependence leads to clusters in the data. The value $\theta < 1$ gives some indication on the clustering behavior and, hence, the dependence in the underlying sequence. $1/\theta$ determines the mean number of exceedances over some threshold per cluster, cf. [2]. Hence, estimates of $1/\theta$ are equal to the ratio of the number of exceedances over the threshold to the number of clusters. Its estimates are only distinguished by different definitions of a cluster in the data.

Regarding the blocks estimate

$$\overline{\theta}^B(u) = \frac{n \sum_{j=1}^{k} \mathbb{1}(M_{(j-1)r,jr} > u)}{rk \sum_{i=1}^{n} \mathbb{1}(X_i > u)} \tag{5}$$

of θ, the cluster is a block of data with at least one exceedance over a threshold u. $M_{i,j} = \max(X_{i+1}, ..., X_j)$, k is the number of blocks, $r = [n/k]$ is the number of observations in each block, and $[\cdot]$ denotes the integer part of a number, cf. [2]. The estimate $\widehat{\theta}$ is selected in a u-region where the plot of the extremal index against u does not change much.



(a) (b)

Fig. 5. Blocks estimates of the extremal index of the maximal Skype PLs within blocks against the number of blocks (a) (cf. [16, Fig. 4a]) and of the IPTV IATs against the threshold u (b)

Example 5: We consider the maximal PLs of a Skype flow within blocks that are separated by those IATs arising from the 98.4% empirical quantile of the IATs, see example 9 in Section 4.2. For the number of blocks equal to 72, $\overline{\theta}^B(u^*)$ is equal to 1, see Fig. 5(a) taken from [16, Fig. 4a]. This implies that the maxima of the PLs corresponding to these 72 blocks are independent. Here $u^* = 197.69$ has been selected which is equal to the mean of the PL maxima.

Let us consider the IPTV IATs. By Fig. 5(b) one can find that $\widehat{\theta} \approx 0.5$ corresponds to the interval of the approximate stability of the plot. It implies the dependence of the IATs.

Estimation of the Hurst Parameter. Fractional Gaussian noise and fractional ARIMA are often used as ideal models of LRD time series. For such models the ACF has the common property $\rho_X(h) \sim c_\rho h^{2(H-1)}$ as $h \to \infty$, and c_ρ is a constant. The value $H = 1/2$ implies $\rho_X(h) = 0$ due to $c_\rho = 0$ and corresponds to independence. The closer the value $1/2 < H < 1$ is to 1 the longer reaches the dependence.

To estimate the Hurst parameter H, we can use the R/S and aggregated variance methods as well as Abry-Veitch's wavelet technique, cf. [1], [23]. These methods assume the self-similarity of the underlying time series. The detection of self-similarity is considered in Section 3.4.

According to the aggregated variance method one plots the logarithm of $\widehat{\mathrm{Var}}X^{(m)}$ (see (1)) versus $\log m$. A straight regression line approximating the points has the slope $\beta = 2H - 2$, $-1 \le \beta < 0$.

According to the R/S method the estimate of the Hurst parameter H is given by a slope of the plot $\log(R(l_i, r)/S(l_i, r))$ against $\log(r)$, where $i = 1, ...K$, and r denotes a range. For this purpose one has to divide the time series into K intervals of length $[n/K]$. $R(l_i, r)/S(l_i, r)$ is computed by the formula

Table 5. Estimation of the Hurst parameter by the data of a Skype flow

r.v.	Estimation methods		
	R/S	Aggregated Variance	Abry-Veitch
Inter-arrival times (sec)	0.6	0.6	0.301 ± 0.023
Packet lengths (bytes)	0.7	0.675	0.729 ± 0.034

$$\frac{R(l,r)}{S(l,r)} = \frac{1}{S(l,r)} \left(\max_{0 \le i \le l} \mu_i(l,r) - \min_{0 \le i \le l} \mu_i(l,r) \right),$$

where

$$S(l,r) = \left(\frac{1}{r} \sum_{i=l+1}^{l+r} \left(X_i - \overline{X_{l,r}} \right)^2 \right)^{1/2},$$

$$\mu_i(l,r) = \sum_{j=1}^{i} \left(X_{l+j} - \overline{X_{l,r}} \right), \qquad \overline{X_{l,r}} = \frac{1}{r} \sum_{i=l+1}^{l+r} X_i,$$

holds for each lag r, starting at points $l_i = i[n/K]+1$ such that $l_i + r \le n$ holds. We may further take the average values of the R/S statistics, $\overline{R(l_i,r)/S(l_i,r)}$, $i = 1, ..., K$.

Following the approach of Abry and Veitch [1] and applying their Matlab code 'LDestimate' [1], one can furthermore compute the wavelet estimate of the Hurst parameter H by means of the slope of a regression line in the logscale diagram, i.e. the log-log plot of the relation between scale $a_j = 2^j$ and the variance estimate $\mu_j = 1/n_j \sum_{k=1}^{n_j} |d_X(j,k)|^2$ of the wavelet details determined by the discrete wavelet-transform coefficients $d_X(j,k)$ of the process X_t.

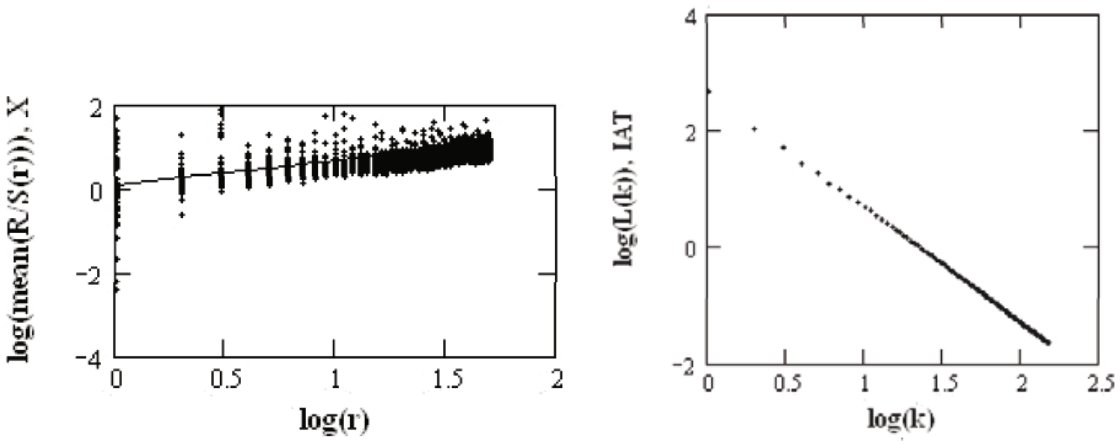
Example 6: We first consider the Skype packet data. The values of the Hurst parameter obtained by the described methods (see Tab. 5, cf. [15]) imply the possibility that no strong LRD occurs regarding the IAT and PL processes of a Skype flow.

Regarding the IATs of an IPTV flow we have obtained $\widehat{H} \approx 0.56$ by the R/S method. It implies a weak long-range dependence of the IATs of an IPTV packet stream, see Fig.6(a), cf. [18].

3.4 Detection of Self-similarity

A time series $\{X_t, t \ge 0\}$ is self-similar or scale invariant with Hurst parameter $H \in (0,1)$ if for any real $a > 0$ and $t \ge 0$ $X_t \equiv^d a^{-H} X_{at}$ holds, i.e. the statistical properties of both sides of this equation are identical.

To check the self-similarity, we apply Higuchi's method, cf. [11]. The latter works as follows. Using a given time series $X_1, X_2, ..., X_n$, one first constructs a new time series X_k^m defined by X_m, X_{m+k}, X_{m+2k},...,$X_{m+[(n-m)/k]k}$, $m = 1, 2, ..., k$. Then one computes



(a) (b)

Fig. 6. Estimation of the Hurst parameter of IPTV IATs by the R/S method (a). Testing the self-similarity by Higuchi's method using $\log \overline{L(k)}$ versus $\log k$ for IPTV IATs (b).

$$L_m(k) = \frac{n-1}{k^2[(n-m)/k]} \sum_{i=1}^{[(n-m)/k]} |X_{m+ik} - X_{m+(i-1)k}|,$$

and draws a log-log plot of the statistic $\overline{L(k)}$ (that is the average value over k sets of $L_m(k)$) versus k. A constant slope D in $\overline{L(k)} \propto k^{-D}$ indicates self-similarity.

Example 7: Considering the IPTV data Fig. 6(b) (see also [18]) shows that the IPTV IATs behave like a self-similar process since the slope of the corresponding plot does not change.

3.5 Application to Packet Data of Skype and IPTV Flows

The results of the statistical characterization of our illustrative data sets obtained in [15] to [18] are summarized in Tab. 6. Apart of the dependent IAT and PL

Table 6. Characterization of Skype and IPTV data (yes = '+', no = '-')

Features	Skype Data			IPTV Data	
	IAT	PL	Block duration	IAT	IAT Block maxima
Independence	−	−	+	−	+
LRD	+	+	−	+	−
Self-similarity	+	+	+	+	+
Heavy-tailed with finite variance	+	−	−	−	+
Heavy-tailed with infinite variance	−	−	+	+	−
Light-tailed	−	+	−	−	−

series, the independent block representatives, namely, the time duration of Skype blocks and the block maxima of IPTV IATs are presented.

4 Principles of Data Segmentation

The statistical analysis of dependent data often requires to partition the data into independent blocks beforehand. Then one can deal with representatives of these blocks in the same manner like with independent data.

For instance, such popular methods like an empirical DF, the maximum likelihood method and many other techniques require i.i.d. data. Following this line of reasoning, our further analysis in Sections 5-7 requires independent blocks, too.

4.1 Equal-Sized Data Blocking

One can simply partition data into equal-sized blocks and find an appropriate minimal block size such that the r.v.s in the blocks are independent and the number of such blocks is large enough. A small number of blocks leads to a small sample size of the representatives of the blocks at our disposal and thus, to a large variance of an estimation by these representatives.

Example 8: We have partitioned the IPTV IATs into equal-sized blocks of the sizes 30, 40, 50, 100, 200, 300, 400, 500, 700, 1000. We have found that 400 is the minimal block size such that the maxima over such blocks are independent. The independence follows from the Ljung-Box test since the test statistic Q does not exceed the 5% quantiles $\chi^2_{0.05}(h)$ of the χ^2 distribution for different lags h, see Tab. 4, cf. also [18].

4.2 Non Equal-Sized Data Blocking

One can partition the time series of packets into non equal-sized blocks that are separated by long time intervals. Then one can expect that the data in the blocks may be independent. More exactly, it is proposed to partition the IATs $X_1, ..., X_n$ into blocks by the exceedances arising from their sufficiently high empirical quantile, see Fig. 7(a), cf. [16]. The exceedances indicate the boundaries of the blocks where the left or the right bound is included in the block. The blocks of the IATs determine the partitioning of the packet sequence and their PLs $Y_1, ..., Y_n$, see Fig. 7(b), cf. [16]. If a partition of the PLs were provided by their own quantile these subsets could be different. The cumulative inter-arrival time lengths within these blocks called block durations are determined by

$$L_j = \sum_{i=k_j}^{k_j-1+N_j} X_i, \qquad j = 1, ..., N_s, \tag{6}$$

cf. [16]. Here N_j is the random size of the jth block depending on the quantile of X_i, $N_0 = 0$, and $k_j = \sum_{m=0}^{j-1} N_m + 1$ is the number of first IAT in the jth block. N_s is the number of blocks.

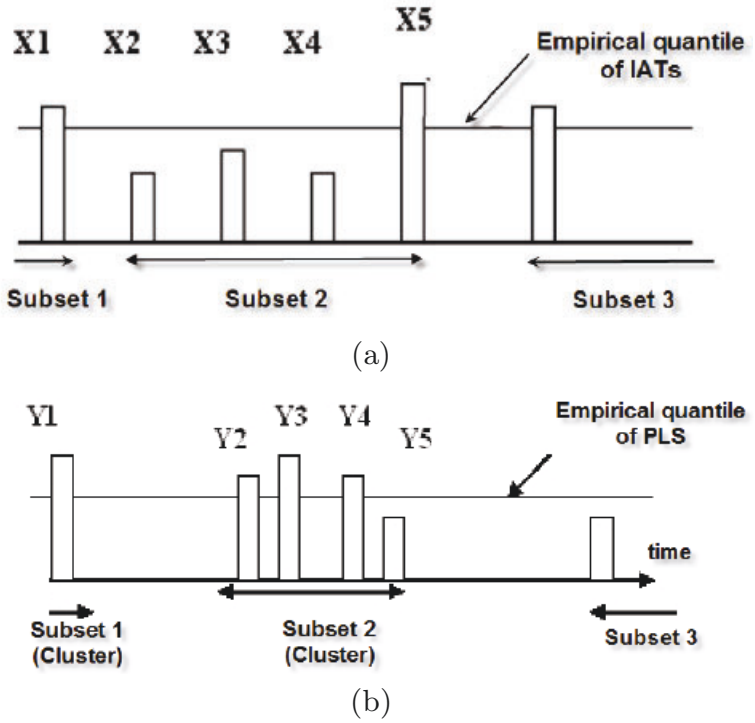


Fig. 7. Partition of the IATs between packets into subsets (blocks) by exceedances over the empirical quantile of the IATs (a). Partition of the PLs into subsets that are separated by long IATs (b).

Example 9: To generate a sufficient number $N_s = 72$ of independent blocks of Skype packets (the description of extremes of these blocks is given in Tab. 1), we use the minimal possible 98.4% empirical quantile of the Skype IATs. It is equal to 0.057 sec. The independence of representatives of such blocks can be easily checked by methods described in Section 3.3. The moderate number of blocks is the price of independence. The characterization of the block durations $\{L_j\}$ is stated in Table 6.

5 Characteristics of Skype User's Satisfaction and Their Estimation

In [8] the bitrate, jitter and round-trip time are selected as factors that influence on the call duration of a Skype session and, hence, on the user's satisfaction. However, the call duration may also reflect the user behavior and cannot be considered as a direct indicator of the user's satisfaction. Here we investigate the IATs between packets received by a Skype client and their PLs which reflect the user's activity and interrelate with the bitrate and delay variation. They influence on the loss and thus on the quality of service (QoS) and quality of experience (QoE) aspects. We propose the mean byte loss, the mean delivery time variation per cluster and the quantiles of lossless periods as indicators of the user's satisfaction.

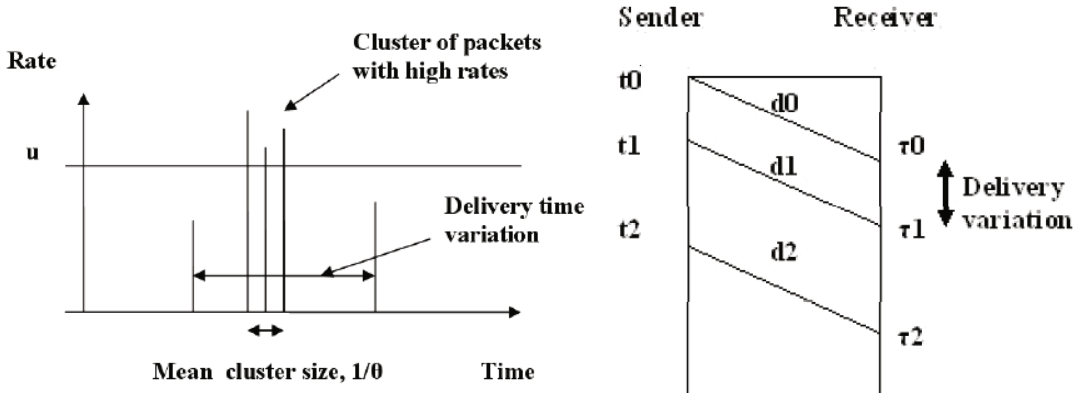


Fig. 8. Clusters of packets are the source of loss (a). Sender-receiver relation of the delivered packets (b).

To analyze all these aspects, we use a bufferless fluid model and assume that the packet stream is approximated by a continuous flow. Its rate is determined by the ratio of the PL Y_i per IAT X_i, i.e. $R_i = Y_i/X_i$, $i = 1, \ldots, n$. It is supposed to be constant between arrivals and only changing at arrival epochs. It is assumed that the IATs are not caused by a silence period of the user but are integral part of one flow.

In case that the IATs between packets are constant, the exceedances of PLs over a threshold u cause loss and delay since the corresponding packets are not delivered. The threshold u is equal to the channel capacity in a time unit. The latter is equal to the IAT. However, in case that the IATs between packets are random, the packets corresponding to exceedances of the required rates $\{R_i\}$ over a capacity u cause loss and delivery delay. Since the rate is defined as ratio of the PL to the IAT, large rates may be generated by frequent packets which arise in the clusters of data, see Fig. 8(a), or by rare large packets. We shall focus on this situation.

The delivery time variation of completely transmitted and correctly received packets is determined by

$$y_i = \tau_i - \tau_{i-1} = t_i + d_i - (t_{i-1} + d_{i-1}) = \Delta t_i + \Delta d_i.$$

Here Δd_i is the delay jitter and Δt_i are the IATs at the sender, see Fig. 8(b). The clusters are generated by packets corresponding to the rates that exceed the channel capacity u. The delivery time variation of packets is equal to the time between two consecutively transmitted and correctly received packets, see Fig. 8(a). Hence, one can estimate the mean delivery time variation of packets per cluster, i.e., the mean IAT between successfully transmitted packets by

$$d = (1 + 1/\theta)\mathbb{E}X.$$

Here $\mathbb{E}X$ is the mean IAT and $1/\theta$ is the mean cluster size, i.e. the mean number of rates exceeding the capacity u per cluster. $1 + 1/\theta$ forms the mean number of

IATs between successfully transmitted packets. θ is calculated by exceedances of the rates $\{R_i\}$ beyond the channel capacity u using the blocks estimator (5). d may be estimated by

$$\widehat{d} = (1 + 1/\overline{\theta}^B(u))\overline{X}$$

where $\overline{X}$ estimates the mean IAT.

Example 10: We consider our Skype flow data where the IATs are random. The estimate of the mean delivery time variation of packets per cluster $\widehat{d} = 0.111$ sec arises for the average IAT $\overline{X} = 0.031$ sec. The extremal index of the rate is given by $\overline{\theta}^B \approx 0.38$ (see Fig. 9) for $k = 150$ equal-sized Skype blocks.

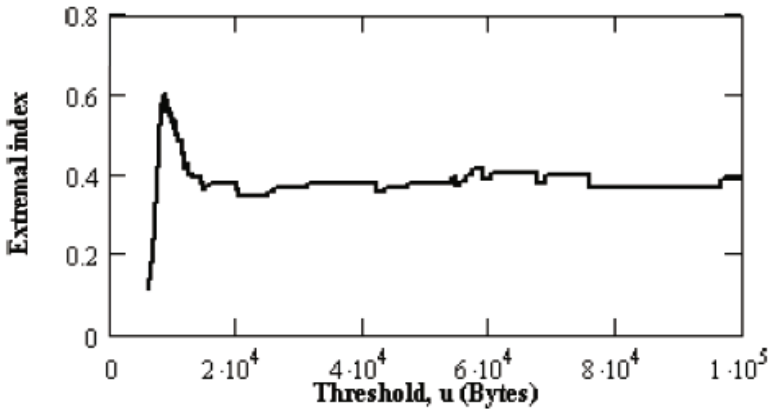


Fig. 9. Blocks estimate of the extremal index of the transmission rate of a Skype flow. The approximate value 0.38 of the extremal index corresponds to the stable u-region and is accepted as an estimate $\overline{\theta}^B$.

Applying a bufferless fluid model, we now estimate the loss, the mean loss and the corresponding channel capacity. First, we consider the case when the IATs are equal. Then the overall byte loss for an observation time

$$E_n(u) = \sum_{i=1}^{n} Y_i \mathbb{1}(Y_i > u)$$

is generated by the PLs $\{Y_i\}$ that exceed a threshold u given in bytes. The latter is equal to the channel capacity in a time unit. The mean byte loss is determined by

$$e_n(u) = \sum_{i=1}^{n} Y_i \mathbb{1}(Y_i > u) / \sum_{i=1}^{n} \mathbb{1}(Y_i > u).$$

Now let us assume that the IATs between packets in the P2P stream are random. Then the loss is generated by packets corresponding to high rates which exceed the channel capacity. Thus, the overall byte loss is given by

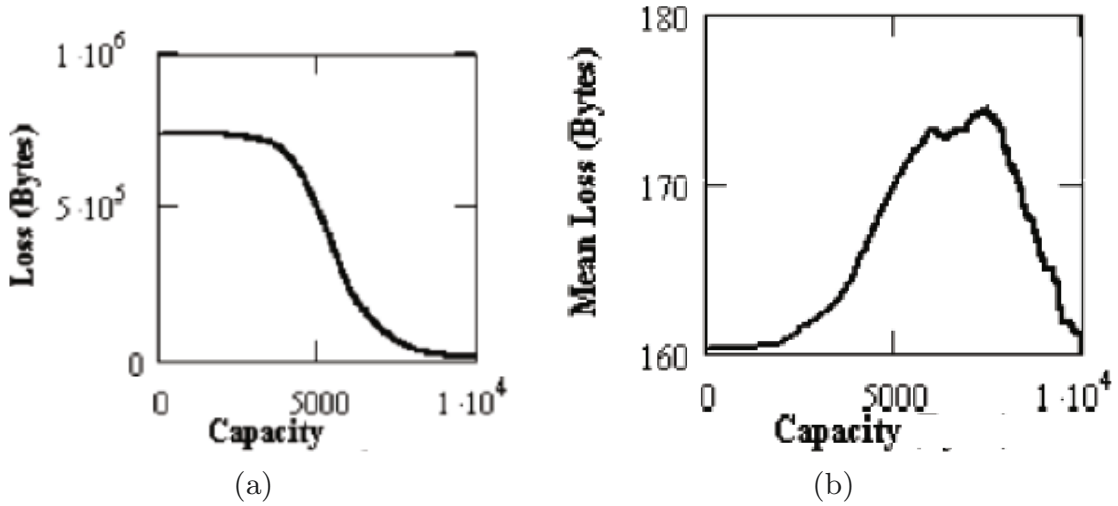
$$\hat{E}(u) = \sum_{i=1}^{n} Y_i \mathbb{1}(R_i > u).$$

The mean byte loss is determined by

$$\hat{e}(u) = \sum_{i=1}^{n} Y_i \mathbb{1}(R_i > u) / \sum_{i=1}^{n} \mathbb{1}(R_i > u), \tag{7}$$

where u denotes the channel capacity.

The time between packets corresponding to rate exceedances beyond the capacity determines a lossless period. The lossless periods may coincide with periods without packet transmission when the rate exceedances correspond to consecutive packets, see 8(a). We calculate the empirical quantiles of lossless periods in the following example.



(a) (b)

Fig. 10. Estimation of the overall byte loss $\hat{E}(c)$ (a) and the mean byte loss $\hat{e}(c)$ (b) against the channel capacity

Example 11: We consider our Skype data with random IATs and calculate $\hat{E}(c)$ and $\hat{e}(c)$, see Fig. 10. The channel capacity corresponding to the 3% overall byte loss is equal to $c^* = 8.534$ kbps. Note that the Internet speech audio codec (iSAC) dynamically adjusts the transmission rate from 10 to 32 kbps. The mean byte loss $\hat{e}(c^*)$ is equal to 168 bytes. $\hat{e}(c)$ increases up to $c_m \approx 7.4$ kbps and decreases beyond c_m. Indeed, the number of packets corresponding to the rates exceeding u (the denominator in (7)) decreases as u increases. However, the numerator of (7) may behave not predictable since large rates can correspond to frequent small packets and frequent (or rare) large packets. The overall byte loss decreases as the capacity increases.

The empirical $50, 75, 80, 85, 95, 97.5, 99.9\%$ quantiles of lossless periods arising from exceedances of the rates $\{R_i\}$ beyond c^* are equal to $30, 34, 35, 36, 39, 44, 121$ ms, respectively. It implies that the loss-free time may exceed these values with the probabilities $50, 25, 20, 15, 5, 2.5, 0.1\%$ and the corresponding overall byte loss is equal to 3%.

6 Estimating the Offered Load

In Section 4 we have considered the partitioning of the packet flow into independent blocks of durations $\{L_j\}$, see (6). Then one can evaluate the overall volume of packets transmitted during a fixed time interval $[0, t]$ by the formula

$$V^*(t) = \sum_{j=1}^{N_t} V_j = \sum_{j=1}^{N_t} \sum_{i=1}^{N_j} Y_i.$$

Here N_t denotes the number of packet blocks arriving before time t, $N_t = \max\{n : t_n < t\}$, and $t_n = \sum_{i=1}^{n} L_i$ is the cumulative time interval corresponding to the n blocks. V_j is the packet volume of the jth block, cf. [17]. For simplicity and without loss of generality, we re-enumerate the packets within blocks from 1 to N_j, where N_j denotes the random number of packets in the jth block.

In [17] it has been shown that the appearances of the volumes $\{V_j\}$ can be considered as a renewal process if we assume that the volume V_j of the jth subset is concentrated at some point of the time interval L_j, e.g. at the beginning. All properties of the renewal process are fulfilled if we assume that the durations of blocks $\{L_j\}$, $j = 1, 2, \ldots$, (or IATs between the cumulative volumes $\{V_j\}$) are i.i.d r.v.s.

To find the expectation of the overall volume at time t, one can use Wald's equation

$$\mathbb{E}(V^*(t)) = \mathbb{E}(\sum_{j=1}^{N_t} V_j) = \mathbb{E}(N_t)\mathbb{E}(V_j) \tag{8}$$

since the volumes of blocks V_j and their number N_t at time t are independent. Then the variance of $V^*(t)$ is determined by

$$\mathrm{Var}(V^*(t)) = \mathrm{Var}(\sum_{j=1}^{N_t} V_j) = \mathrm{Var}(V_j)\mathbb{E}(N_t) + (\mathbb{E}(V_j))^2 \,\mathrm{Var}(N_t) \tag{9}$$

if the second moment of the volume exists, i.e. $\mathbb{E}V_j^2 < \infty$ cf. [14], [25].

Furthermore,

$$H(t) = \mathbb{E}(N_t) = \sum_{n=1}^{\infty} \mathbb{P}\{t_n < t\}$$

denotes the renewal function, cf. [14]. It exhibits simple analytic forms just for a few distributions of the IATs like the uniform, normal and exponential distributions, e.g., it is linear $H(t) = \lambda t$ for an exponential distribution with intensity λ. Usually, the distribution of the IATs $\{L_i\}$ of packet volumes V_j is unknown. Thus, one has to apply nonparametric estimators of $H(t)$, e.g., a histogram-type estimator recommended in [14]. Using the IATs $\{\tau_j, j = 1, \ldots, N\}$, it is determined by

$$\widetilde{H}\,(t, k, N) = \sum_{n=1}^{k} \frac{1}{N_n} \sum_{i=1}^{N_n} \mathbb{1}(t \geq t_n^i).$$

Here $t_n^i = \sum_{q=1+n(i-1)}^{n\cdot i} \tau_q$, $i = 1, \ldots, N_n$, and $N_n = \left[\frac{N}{n}\right]$, $n = 1, \ldots, k$, are observations of the r.v. t_n. In our consideration we take $\{L_j\}$ and N_S instead of $\{\tau_j\}$ and N. N_S denotes the number of blocks or the number of $\{L_j\}$, respectively.

To select the parameter k for a fixed t one can use the formula

$$k^* = \arg\min\{k : \widetilde{H}(t,k,N) = \widetilde{H}(t,k+1,N), k = 1, \ldots, N-1\}.$$

More details about these methods can be found in [14, Chapter 8].

Due to the limited number of IATs $\{L_i\}$ that are at our disposal the histogram-type estimate becomes constant after a sufficiently large time t, [1] i.e., $\widetilde{H}(t,k,N_S) = k$ holds for $t \in [t_{\max}(k), \infty)$, where

$$t_{\max}(k) = \max_{1\le n\le k} \max_{1\le i\le [N_S/n]} t_n^i \le \sum_{i=1}^{N_S} L_i$$

and $k \le N_S$ is some fixed number.

For large t one can use the well-known linear approximation of the renewal function

$$H(t) = \frac{t}{\mu} + \frac{\sigma^2}{2\mu^2} - \frac{1}{2} + o(1),$$

if the mean μ and variance σ^2 of $\{L_j\}$ are finite[2]. Another approximation

$$H(t) = \frac{t}{\mu} + \frac{t^2\,(1-F(t))}{\mu^2(\alpha-1)(2-\alpha)} + o(1) \tag{10}$$

for $t \to \infty$ (cf. [24]), where $F(t)$ is the DF of L_j, is valid for regularly varying distributions (2) and $1 < \alpha < 2$. This is the case if the variance of $\{L_j\}$ is infinite. One can rewrite (10) using the estimate $t^{-\alpha}$ instead of $1 - F(t)$ and replacing μ and α by their estimates, e.g., by $\widehat{\mu}$ which is the average of $\{L_j\}$ and by Hill's estimate $\widehat{\alpha} = 1/\widehat{\gamma}^H\,(n, k_0)$, respectively. Then we get

$$\widehat{H}(t) = \frac{t}{\widehat{\mu}} + \frac{t^{2-\widehat{\alpha}}}{\widehat{\mu}^2(\widehat{\alpha}-1)(2-\widehat{\alpha})} \tag{11}$$

for $t > t_{\max}(k)$.

According to [17] one can estimate the mean overall volume $\mathbb{E}(V^*(t))$ at time t by the formula

$$\overline{V^*}(t) = H^*(t)\overline{V}. \tag{12}$$

Here $\overline{V}$ is the sample average of $\{V_j\}$ which can be used instead of $\mathbb{E}(V_i)$ in (8),

$$H^*(t) = \begin{cases} \widetilde{H}\,(t,k,N_S)\,, & t \le t_{max}, \\ t/\widehat{\mu} + \widehat{\sigma}^2/(2\widehat{\mu}^2) - 1/2, & t > t_{max}, \sigma^2 < \infty, \\ \widehat{H}(t), & t > t_{max}, \sigma^2 = \infty, \end{cases} \tag{13}$$

and $\widehat{\sigma}$ is the standard deviation of $\{L_j\}$.

[1] This is a typical feature of all histogram-type estimates.

[2] The notation $o(1)$ means that the approximation is valid up to an arbitrary constant.

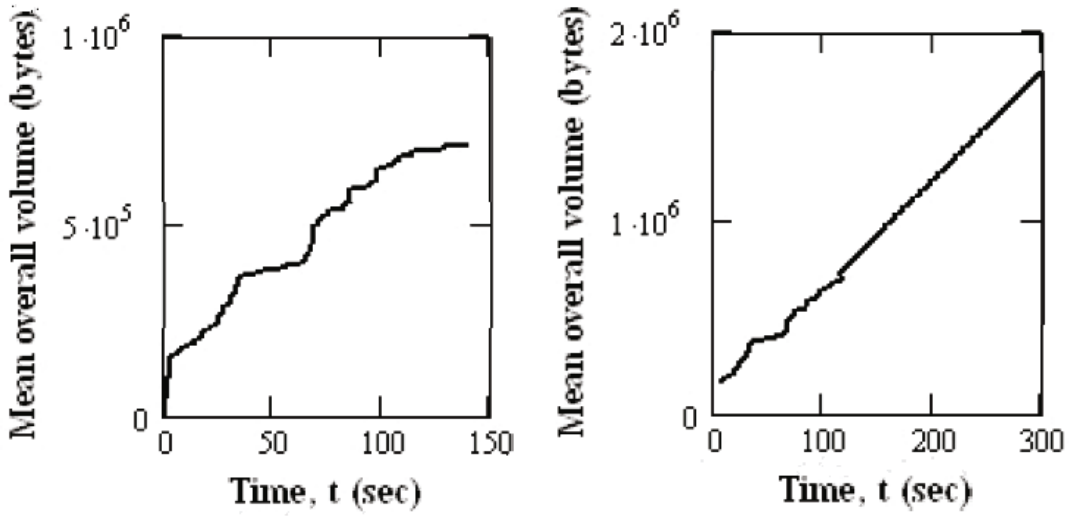


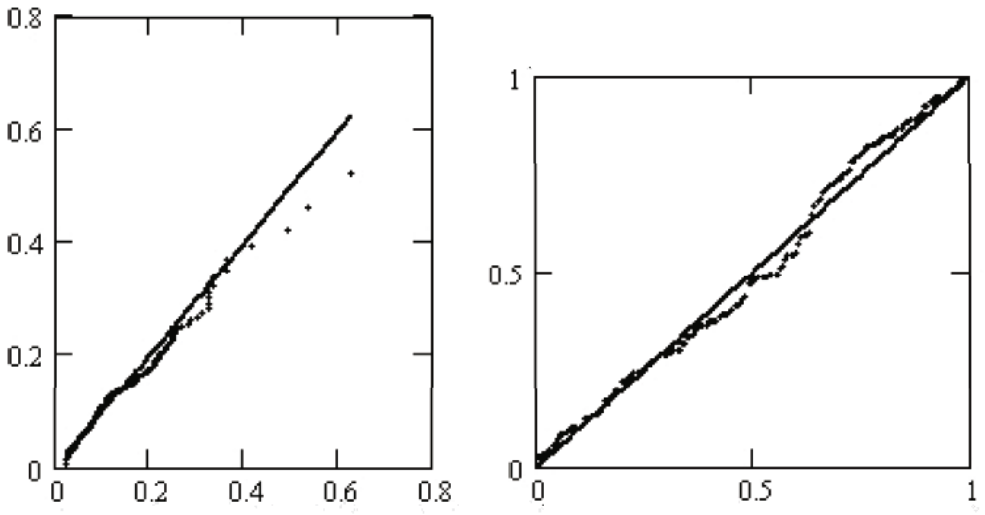
Fig. 11. The estimate of the mean overall volume $\overline{V^*}(t)$ in the time intervals $[0, t_{max}] = [0, 138.914]$ (a) and $t \in [5, 300]$ (b). For $t > 138.914$ $\widetilde{H}\,(t, k, N) = k^* = 72$ holds and it is replaced by a linear model (cf. [17, Fig. 7])

Example 12: We consider again the Skype data. In this case the variance of the block volume $\mathrm{Var}(V_j)$ is infinite, since the tail index of V_j is about 1.5 as shown in [17]. Then the variance $\mathrm{Var}(V^*(t))$ is infinite which follows from (9). Thus, we can estimate $\mathbb{E}(V^*(t))$ by formulae (12) and (13) only. The sample average of $\{V_j\}$ is given by $\overline{V} = 10.18$ Kbytes. The Hill's estimate of the tail index α of $\{L_j\}$ falls into the interval $[1.468, 1.56]$ and one can expect that the distribution of $\{L_j\}$ is regularly varying, cf. [17]. Hence, we can apply (11) and take $\alpha = 1.5$. Figures 11(a), 11(b) (see also [17, Fig. 7]) depict the mean offered load of the packets corresponding to a pre-defined time t. In Figure 11(b) the estimate of $H(t)$ coincides with $\widetilde{H}\,(t, k, N_S)$ in $t \in [5, 138.914]$ and with (11), where $\alpha = 1.5$ holds for $t \in (138.914, 300]$. The mean overall volume increase evidently as time increases. One can calculate how much traffic load arises at time t by means of $\overline{V^*}(t)$. The considered Skype traffic is non-Poissonian, since $H^*(t)$ and $\overline{V^*}(t)$ are not linear at relatively small times t. For Poisson traffic $H^*(t) = \lambda t$ holds and $\overline{V^*}(t)$ is linear at any t.

7 Distribution of the Maximum of Inter-Arrival Times

In Section 4 we have considered the partitioning of the packet flow into independent blocks. Knowing these blocks, one can fit the distribution of representatives of these blocks accurately.

We consider here the IPTV data and their equal-sized blocks of size 400 described in the example 8 of Section 4.1. Then we can fit the distribution of the block maxima $\{X_i^{400}\}$. It is well known that the Generalized Extreme Value (GEV) distribution with DF



(a) (b)

Fig. 12. The QQ- and PP-plots (a) and (b), respectively, of the GEV distribution of the IPTV IAT block maxima with parameters $\gamma = 0.21666$, $\sigma = 0.05887$, $\mu = 0.0881$.

$$F(x) = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma}\right)$$

is an appropriate model to fit the maximum. Since the block maxima $\{X_i^{400}\}$ are independent, we can apply the maximum likelihood method to find the parameters of a GEV. By different goodness-of-fit tests with a 5% confidence level the following values $\gamma = 0.21666$, $\sigma = 0.05887$ and $\mu = 0.0881$ were found to be the best ones, see Tab. 7. They provide QQ- and PP-plots which are close to the empirical data, see Fig. 12 (cf. also [18]).

Table 7. Test results of the GEV distribution fitted to IAT block maxima $\{X_i^{400}\}$

Parameters	**Goodness-of-Fit Tests**	
	Kolmogorov-Smirnov	**Anderson-Darling**
$\gamma = 0.21666$, $\sigma = 0.05887$, $\mu = 0.0881$	0.07075	0.8341

Using (4) now and a GEV approximation of $\mathbb{P}\{\widetilde{M}_n \le u\} = \mathbb{P}\{X_1^{400} \le u\}$, we can approximate the distribution of the maximum of the IPTV IATs by the formula

$$\mathbb{P}\{M_n \le x\} \approx \exp\left(-\left(1+\gamma\frac{x-\mu^*}{\sigma^*}\right)^{-1/\gamma}\right). \tag{14}$$

Table 8. The GEV distribution fitted to the IAT maximum and its high quantiles

Parameters	**High quantiles**		
	95%	97.5%	99%
$\gamma = 0.21666,\ \sigma^* = 0.05,\ \mu^* = 0.051$	0.263	0.337	0.452

Here $\mu^* = \mu - \sigma(1-\theta^\gamma)/\gamma$, and $\sigma^* = \sigma\theta^\gamma$ hold, cf. [2, p. 377]. We can also find the quantiles of this maximum. In Fig. 5(b) we have shown that $\widehat{\theta} \approx 0.5$ arises for the IPTV IATs. Then one can calculate the new parameters μ^* and σ^* by this estimate $\widehat{\theta}$. Regarding the IPTV IAT the quantiles of the maximum M_n can be obtained by the formula

$$q_p = \frac{(-\ln(1-p))^{-\gamma} - 1}{\gamma}\sigma^* + \mu^*,$$

see Table 8, cf. [18]. The high $95, 97.5, 99, 99.9\%$ quantiles imply a delay between packets which can only be exceeded with the small probabilities $5, 2.5, 1, 0.1\%$, respectively.

8 Conclusions

In recent years peer-to-peer (P2P) multimedia applications have became a powerful service platform for the generation and transport of voice and video streams over IP. The application of variable bitrate encoding schemes and the packet based voice and video transfer raises a large variety of new questions regarding traffic characterization.

In our study we have developed a general methodology concerning the statistical analysis of P2P packet flows using two types of information. These characteristics are given by the inter-arrival times between packets and the lengths of the transported packets. We have focussed on the case when the inter-arrival times are random entities. Such a situation arises, for instance, when a wireless access to the Internet by Skype or IPTV clients is considered.

We deal with observations which are time series, i.e. they are dependent. Thus, we have presented principles of data blocking to partition the observations into independent blocks and to deal with independent data instead of dependent ones.

Our methodology includes the statistical characterization of a P2P packet stream regarding the stationarity, long-range dependence, self-similarity and heaviness of tail of the associated distributions. In addition to that, we have presented some important characteristics of the Skype user's satisfaction such as the overall byte loss, the mean byte loss, the mean delivery time variation of packets per cluster and the quantiles of lossless periods. These characteristics extend the list of known indices like the bitrate, jitter and round-trip time.

The proposed statistical methodology is accompanied by several examples which illustrate its application to packet flows of representative Skype and IPTV sessions.

Further, we have considered the problem to evaluate the channel capacity which is required to guarantee an appropriate overall byte loss. We have always assumed that the loss is caused by packets corresponding to exceedances of the transmission rate beyond the channel capacity of a bufferless fluid model and that the rate is equal to the ratio of the packet length to the adjacent inter-arrival time. This is a natural assumption for random inter-arrival times between packets.

Moreover, we have estimated the offered traffic load in a finite observation period. For this purpose we have observed that the cumulative traffic volumes arising from independent blocks of packets create a renewal process.

In summary, the proposed statistical methodology provides a powerful and versatile approach for a traffic characterization in the Internet. It can be applied to any correlated data arising from monitored packet streams and is not limited to P2P data used here as illustration.

Acknowledgment. The authors acknowledge the partial support by research grants of the EU FP6-NoE project "EuroFGI" under contract 028022, the COST Action IC0703 "Data Traffic Monitoring and Analysis (TMA)", and the German Ministry of Education and Reseach (BMBF) under contract MDA 08/015.

Regarding the measurement studies of P2P traffic the authors are also grateful to Mr. Schweßinger and Mr. Eittenberger for their support.

References

1. Abry, P., Veitch, D.: Wavelet Analysis of Long-Range Dependence Traffic. IEEE Transactions on Information Theory 44(1), 2–15 (1998)
2. Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
3. Bendat, J.S., Piersol, A.G.: Random Data: Analysis and Measurement Procedures. J. Wiley & Sons, New York (1986)
4. Beran, J.: Statistics for Long-Memory Processes. Chapman & Hall, New York (1994)
5. Bhattacharya, R.N., Gupta, V.K., Waymire, E.: The Hurst effect under trends. J. Appl. Probab. 20, 649–662 (1983)
6. Bonfiglio, D., Mellia, M., Meo, M., Rossi, D., Tofanelli, P.: Revealing Skype Traffic: When Randomness Plays with you. In: Proceedings of ACM SIGCOMM 2007, Kyoto, August 27–31 (2007)
7. Brockwell, P.J., Davis, R.A.: Introduction to Time Series and Forecasting, 2nd edn. Springer Texts in Statistics, New York (2002)
8. Chen, K.-T., Huang, C.-Y., Huang, P., Lei, C.-L.: Quantifying Skype user satisfaction. In: Proceedings ACM SIGCOMM 2006, Pisa, Italy, September 11-15 (2006)
9. Cochran, W.G.: The distribution of the largest of a set of estimated variances as a fraction of their total. Ann. of Eugenics 11, 47–52 (1941)
10. Giraitis, L., Leipus, R., Philippe, A.: A test for stationarity versus trends and unit roots for a wide class of dependent errors. Econometric Theory 22(6), 989–1029 (2006)

11. Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Physica D 31, 277–283 (1988)
12. Liu, F., Li, Z.: A Measurement and Modeling Study of P2P IPTV Applications. In: Proceedings of the 2008 International Conference on Computational Intelligence and Security, vol. 1, pp. 114–119 (2008)
13. Ljung, G.M., Box, G.E.P.: On a Measure of Lack of Fit in Time Series Models. Biometrika 65, 297–303 (1978)
14. Markovich, N.M.: Nonparametric Estimation of Univariate Heavy-Tailed Data. J. Wiley & Sons, Chichester (2007)
15. Markovich, N.M., Krieger, U.R.: Statistical Analysis of VoIP Flows Generated by Skype Users. In: Proceedings of IEEE International Workshop on Traffic Management and Traffic Engineering for the Future Internet, FITraMEn, Porto, Portugal, December 11-12 (2008)
16. Markovich, N.M., Krieger, U.R.: Statistical Characterization of QoS Aspects Arising From the Transport of Skype VoIP Flows. In: Proceedings of The First International Conference on Evolving Internet (INTERNET 2009), IARA, Cannes/La Bocca, August 23-29, pp. 9–14 (2009)
17. Markovich, N.M., Krieger, U.R.: Statistical Analysis and Modeling of Skype VoIP Flows. Computer Communications 33, 11–21 (2010)
18. Markovich, N.M., Krieger, U.R.: Characterizing Packet Traffic in Peer-to-Peer Video Applications. Technical Report, Otto-Friedrich University Bamberg (2009) (submitted)
19. Novak, S.Y.: Inference of heavy tails from dependent data. Siberian Advances in Mathematics 12(2), 73–96 (2002)
20. Resnick, S.I.: Heavy-Tail Phenomena. Probabilistic and Statistical Modeling. Springer, New York (2006)
21. Runde, R.: The asymptotic null distribution of the Box-Pierce Q-statistic for random variables with infinite variance. J. of Econometrics 78, 205–216 (1997)
22. Silverston, T., Fourmaux, O., Botta, A., Dainotti, A., Pescapé, A., Ventre, G., Salamatian, K.: Traffic analysis of peer-to-peer IPTV communities. Computer Networks 53, 470–484 (2009)
23. Taqqu, M.S., Teverovsky, V., Willinger, W.: Estimators for long-range dependence: an empirical study. Fractals 3, 785–798 (1995)
24. Teugels, J.L.: Renewal theorems when the first or the second moment is infinite. Annals of Statistics 39, 1210–1219 (1968)
25. Trivedi, K.S.: Probability & Statistics with Reliability, Queuing, and Computer Science Applications. Prentice Hall of India, New Delhi (1997)
26. Skype, `http://www.skype.com`
27. SopCast, `http://www.sopcast.com`