Analyzing Measurements from Data with Underlying Dependences and Heavy-Tailed Distributions

Natalia M. Markovich
Institute of Control Sciences
Russian Academy of Sciences
117997 Moscow, Russia
markovic@ipu.rssi.ru

Udo R. Krieger
Computer Networks Group
Otto-Friedrich University
D-96052 Bamberg, Germany
udo.krieger@ieee.org

ABSTRACT

We consider measurements that are arising from a next generation network and present advanced mathematical techniques to cope with the analysis and modeling of the gathered data. These statistical techniques are required to study important performance indices of new real-time services in a multimedia Internet such as the demanded bandwidth or delay-loss profiles of packet flows during a session. The latter data sets incorporate strongly correlated or long-range dependent time series and heavy-tailed marginal distributions determining the underlying random variables of the data features. To illustrate the proposed statistical analysis concept, we use traces arising from the popular peer-to-peer video streaming application SopCast.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement techniques, modeling techniques

General Terms

Performance

Keywords

Data analysis, heavy-tailed distributions, long-range dependence, NGN traffic characterization, peer-to-peer packet traffic

1. INTRODUCTION

Nowadays, peer-to-peer (P2P) video and live TV applications generate a rapidly growing proportion of the overall Internet packet traffic. In the current setting these real-time applications can use an overlay network. It is established among the corresponding clients on the hosts of the users to disseminate the shared multimedia information in a swarm like fashion by means of an underlying network infrastructure at its bottom (see Fig. 1). The latter employs the TCP/IP protocol stack with certain quality-of-service (QoS) enhancements.

The multimedia information such as MPEG-4 encoded video files

ICPE'11, March 14–16, 2011, Karlsruhe, Germany.

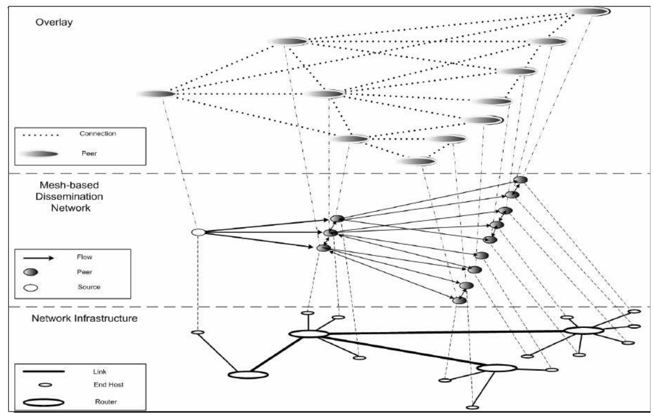


Figure 1: Peer-to-peer overlay network

is normally decomposed into fixed sized pieces called chunks, dynamically send into the overlay network by the source of the original multimedia object and then continuously disseminated piecewise among the peer population based on index information.

The clients of the overlay called peers can retrieve this multimedia information by means of these chunks from other peers contributing to the generated overlay network of a certain group of clients that is sharing the same multimedia file and additionally offer their stored pieces of a demanded object. In the relevant cases studied here, the P2P media sharing protocol establishes a full mesh at the overlay, provides means to identify the clients in the overlay by a tracker service and employs a pull protocol to manage the access of a peer to the requested content stored at the neighbors. Thus, the peers combine client and server roles on an equal basis. A peer which has collected all pieces of a certain segment of a real-time media stream without any gaps will assemble the latter into the original media frames and then start to deliver it by its streaming engine to a user observing the streamed multimedia content.

Gaps in the stream must be filled in advance requesting missing pieces of the stream from the peer population by appropriate signaling messages. Hence, a composite stream of multimedia content and signaling packets is incoming to and outgoing from a peer in the overlay network and traversing its underlying access link.

The key issue is that peers in such a P2P overlay network may randomly join or leave the overlay structure. Then leaving peers cause a loss of all their stored multimedia data and, hence, downgrade the quality of experience of a user watching the multimedia stream (cf. [5]). The problem is known as churn issue of P2P networks.

It is the objective of our statistical investigations to provide a general, solid mathematical methodology to define, analyze and estimate important quality indices of such packet traffic carrying real-time information over a P2P network. For this purpose we have to control the loss and delay profiles of the information flows at the packet layer, but not at the peer layer (cf. [5]). If multimedia traffic from certain peers that is requested by an interested peer is missing or heavily delayed, it cannot contribute to the aggregated flow feeding the streaming engine of a user. Therefore, the latter has to send the same object requests to other peers which generates an additional delay. This operation implies that we must propose an evaluation method regarding the losses and delays of the packet flows of a shared multimedia stream that are transmitted to an observing peer along the network infrastructure.

Considering the control of the transmission quality, it is the objective of our study to propose the delay of a successful packet delivery along a bottleneck link, the lossless time of a packet transmission and the corresponding byte loss as new quality indices of a continuous stream of multimedia packets. These three indices are studied since they substantially impact on the visual interpretation of a live stream. In this regard we do not take into account the failure rate of the peers as a source of loss (cf. [5]).

The analysis of the marginal distributions and means of these indices and the estimation of their corresponding quantiles are proposed here. High quantiles close to 100% reveal approximate upper bounds on the quality indices of a flow. The quality of the packet transmission process is good if the high quantiles of the delay are relatively small. High quantiles of the lossless (or burstless) periods indicate those values which can only be exceeded with a negligible small probability. Relatively large lower quantiles of the lossless periods may imply a good quality of the transmission process. Furthermore, quantiles of both the delays and lossless periods can be compared with the mean or maximum of the inter-arrival times (IATs) between packets to assess the stability of the transmission process.

To achieve these QoS assessment goals, we consider statistical problems arising in P2P packet transmission like the dependence and heaviness of the tail of the underlying distributions of the quality indices. These issues may seriously disturb the estimation of the quality indices. To compute the new indices, we only analyze the corresponding IATs of packets, the associated packet lengths (PLs) and corresponding instantaneous transmission rates of a packet flow towards a single representative peer of the overlay.

To illustrate our concepts, we study the aggregated packet flow of a P2PTV application that is exchanged by a selected peer on a wireless bidirectional link. It is generated by the popular dissemination platform SopCast. In our study we do not aim to compare the features of other important P2PTV applications.

Measurement studies of relevant P2PTV applications like PPlive, PPStream, SopCast, and TVants have already been performed, for instance by [6, 14, 16] and [17] among others. In [14] the traffic properties and peer behavior of each application has been considered at the packet level distinguishing the direction of the transmission and the signaling and video traffic. The traffic directions are shown to be similar with regard to energy spectra for all four applications (cf. [14]). The descriptive statistics of the IATs between packets, the packet size and rate, the peer lifetime and the throughput have been used as major indices of the traffic characteristics. In [6] the behavior of P2PTV applications is considered at the flow level where peer nodes act as content servers. The IATs between flows, the flow size, as well as the arrival rate and the duration of the flows were selected as major characteristics. It has been shown that the distributions of these indices belong to the class of heavy-tailed distributions. In [16] the operational mechanism of SopCast has been investigated.

In our study we use an approach based on Wald's equation, a high quantile estimation (i.e. a quantile close to 100%) and the exceedances of the instantaneous transmission rates of a composite P2P flow of multimedia and signaling packets over a threshold to estimate the means and quantiles of the delay of a successful packet delivery to a peer, the associated byte loss and the lossless time. Great attention is devoted to demonstrate the proposed statistical methodology by examples of real SopCast flows.

We have proved in [10] that the IAT sequence of a live TV application generated by a SopCast client constitutes a weak long-range dependent (LRD), self-similar and heavy-tailed stochastic process with infinite variance and finite mean. This LRD property of SopCast flows that impacts the QoS parameters has also been confirmed by [14].

In contrast to [14] we explain here the mechanism of this impact. We use a fluid model of a bufferless logical transmission channel of the packet streams. Then we assume that the required instantaneous transmission rate of a packet flow can be approximated by the ratios $R_i = Y_i/X_i$, $i = \overline{1,n}$, where $\{X_i, i \in \mathbb{N}_n\}$, $\mathbb{N}_n = \{i \in \mathbb{N} | i \leq n\}$, $\{Y_i, i \in \mathbb{N}_n\}$ and $\{R_i, i \in \mathbb{N}_n\}$ denote the IATs between packets, the PLs and the rates of transmission. Roughly speaking, the dependence of the IATs causes clusters of packets whose arrival times are close to each other. Due to relatively small IATs within clusters the demanded transmission rates of the corresponding packets may exceed the available capacity of the transmission channel at a bottleneck link. This behavior generates a loss and delay regarding the successful packet delivery which substantially impacts on QoS. Considering the corresponding packet stream with its variable bitrate requirement, the latter rates determine in this way its equivalent bandwidth subject to these QoS constraints. The weak dependence and a finite mean of the IATs allow us to estimate the expectation of the IAT by the sample mean[1] and hence, to estimate the mean lossless time and the mean delay.

The quality indices, i.e. the delay of a successful packet delivery, the lossless time and the byte loss, may be dependent random variables (r.v.s) for a sufficiently low available channel capacity. Therefore, we propose a so called declustering of the flow data to estimate the quantiles of these quality indices. The reason is that the estimators of the quantiles require independent r.v.s. A declustering implies the separation of the observations of the underlying random indices into independent blocks. Then we can select the maxima within these blocks called block maxima as new representatives of our data set. Further, we estimate the quantiles of the delay, byte loss and lossless time by means of the distributions of their block maxima that behave statistically independent. It has been shown in [10] that the distribution of the maxima of the IATs of a SopCast flow obeys a Generalized Extreme Value (GEV) distribution (cf. [12]).

The delay of a packet delivery may not only be caused by the loss of packets but also by large IATs between packets which are related to a silence period of the source or a peer failure. Therefore, the study of the distribution of the sample maximum of the packet IATs and its quantiles may be important, since large delays in the packet delivery of video chunks may severely reflect on the visual interpretation of a live stream.

In the paper we also investigate the impact of the signaling traffic on the proposed quality indices and compare these indices both for a composite traffic flow which includes both signaling packets and video content and the pure video traffic of a SopCast session.

The paper is organized as follows. In Section 2 we describe the

[1]This follows from the law of large numbers.

measurement setting of the P2P SopCast video flows. In Section 3 the signaling packets are excluded from the underlying composite SopCast flow to investigate the pure video traffic. In Section 4 the estimation of the means and quantiles of the quality indices, i.e. the delay, the byte loss and the lossless time, is formulated and the motivation of further packet traffic characterization is explained. The rest of the paper contains the characterization and the estimation of the quality indices of monitored SopCast video flows in Section 5. In particular, the high quantiles of the quality indices are estimated. The means of the delay, the byte loss and the lossless time of the total traffic are estimated by the underlying P2P data in Section 5.2. Finally, the findings are summarized in the conclusions.

2. MEASUREMENT SETTING

To collect appropriate P2PTV traces of SopCast sessions, a measurement study has been performed by the Computer Networks Laboratory at Otto-Friedrich University Bamberg, Germany, during the second quarter of 2009. It has focused on a typical scenario of a wireless access to the Internet.
The test bed has been developed to study the basic operation of the P2PTV streaming system SopCast in a representative home environment of a subscriber (cf. Fig. 2). In this wireless scenario a

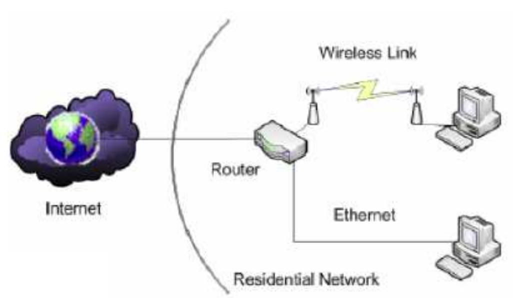
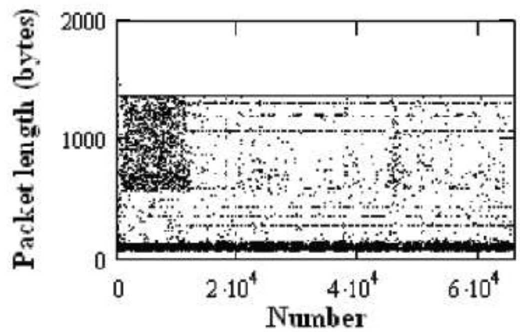


Figure 2: Wireless test bed

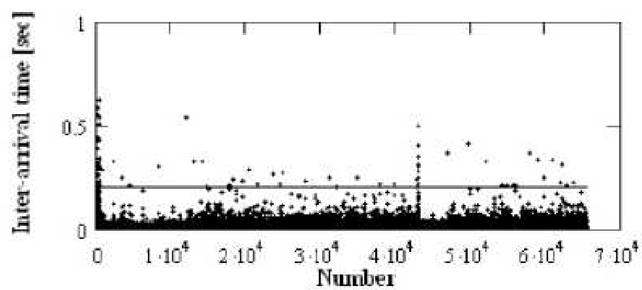
SopCast client is running on a desktop PC IBM Thinkcentre with 2.8 GHz Intel Pentium 4 processor, 512 MB RAM under Windows XP Home. It is attached by a Netgear WG111 NIC operating the IEEE802.11g MAC protocol over a wireless link to the corresponding ADSL router acting as gateway to the Internet.
Watching a popular sport channel, representative traces arising from sessions of 30 minutes have been gathered by Wireshark on the data link layer at the portable host. Since it is our main objective to propose a common methodology for the proper estimation of important quality indices regarding the packet transmission of a multimedia stream, the possibly low accuracy of the IATs recorded by a Wireshark agent does not constitute principle problems.
Subsequently, we shall use these gathered P2PTV traces to illustrate our mathematical approach and all related difficulties of the applied statistical analysis.

3. COMPOSITE P2PTV TRAFFIC AND ITS VIDEO CONTENT

We consider a packet trace of the superimposed flows exchanged with a single observed peer in the overlay network during a SopCast session (cf. Fig. 1). The recorded composite packet traffic that is incoming to and outgoing from the peer uses UDP as transport protocol and contains both large-sized, frequent packets with small IATs in terms of packet clusters corresponding to the exchange of video information and small-sized, rare packets with large IATs carrying signaling information (cf. [14], [16]). It has been pointed out by [14] that SopCast produces more signaling traffic than other P2P systems such as PPLive, PPStream and TVAnts. Signaling packets have fixed PLs that correspond to horizontal lines in the scatter plot shown in Fig. 3(a). It can be expected that they instantiate PLs which are less than or equal to 130 bytes UDP payload.
To investigate the mean delay of a successful packet delivery to the observed peer and the mean lossless time of pure video traffic, we exclude from this composite traffic those packets whose lengths are less than 130 bytes. In [16] the lengths of SopCast signaling packets are indicated by a UDP payload of $\{28, 42, 46, 52, 80\}$ bytes and video content by packets of $\{377, 497, 617, 1081, 1201, 1320\}$ bytes. In [14] the joint probability distribution of the PLs and the IATs of the mentioned P2PTV applications has revealed two main clusters of packets: small packets whose lengths is less than 200 bytes with large IATs and large packets whose lengths is larger than 1000 bytes with small IATs. The scatter plot of the video packets of our illustrative SopCast session shows that there are three classes of video packets with relatively large IATs, namely, one class with PLs larger than 1000 bytes and two classes with PLs centered approximately at 600 bytes and at 180 bytes (see Fig. 3(a)). There are many frequent packets of different size larger than 130 bytes, too, which is in agreement with [16].
The virtual exclusion of the signaling traffic which is located mostly between clusters of video packets may increase the IATs of these video packets corresponding to the inter-cluster times (see Fig. 4(a)). This can lead to large video packets with virtually larger IATs and smaller rates than those ones related to the total traffic, see Table 1.



(a)



(b)

Figure 3: (a) Scatter plots of the packet lengths of the composite SopCast traffic; (b) IATs between packets of the aggregated flow to one peer together with their 99.9% empirical quantile corresponding to 0.2 (solid horizontal line).

Table 1: Description of the IATs between packets, PLs and transmission rates arising from the aggregated flow to an observed peer.

R.V.	Type of Traffic	Sample Size	Min	Max	Median	Mean	StDev	Skew-ness	Kurtosis
IAT	Total traffic	$6.553 \cdot 10^4$	$2.1 \cdot 10^{-5}$	0.625	$5.58 \cdot 10^{-4}$	$4.934 \cdot 10^{-3}$	0.016	13.56	313.087
[sec]	Video traffic	$1.798 \cdot 10^4$	$2.6 \cdot 10^{-5}$	2.691	$3 \cdot 10^{-3}$	0.018	0.072	10.623	199.478
PL	Total traffic	$6.553 \cdot 10^4$	60	1506	84	400.308	545.451	1.14	-0.649
[bytes]	Video traffic	$1.798 \cdot 10^4$	132	1506	1362	1256	274.048	-2.679	6.171
Rate$\cdot 10^{-5}$	Total traffic	$6.553 \cdot 10^4$	$9.601 \cdot 10^{-4}$	544.8	1.763	12.616	30.543	8.845	115.633
[bps]	Video traffic	$1.798 \cdot 10^4$	$7.129 \cdot 10^{-4}$	523.8	4.194	10.407	42.21	9.368	91.849

Looking, for example, at Fig. 4(a), the exclusion of the expected "signaling" packets with the numbers 6 and 7 increases the IAT between the packets 5 and 8.

The non-zero skewness and kurtosis of the IATs (see Tab. 1) imply a deviation from the normal distribution. Further the mean of the IATs is larger than the median. It indicates a long tail of the corresponding distribution of the IATs. The numerous exceedances over the empirical quantile at 99.9% in Fig.3(b) indicates that this distribution may be heavy tailed. The heaviness of the tail of the IATs is derived by a calculation of the tail index and the mean excess function (cf. [10], see subsection 4.2.3).

4. STATISTICAL DATA ANALYSIS AND ESTIMATION OF THE QUALITY INDICES

4.1 Packet Delivery Delay, Lossless Time and Byte Loss

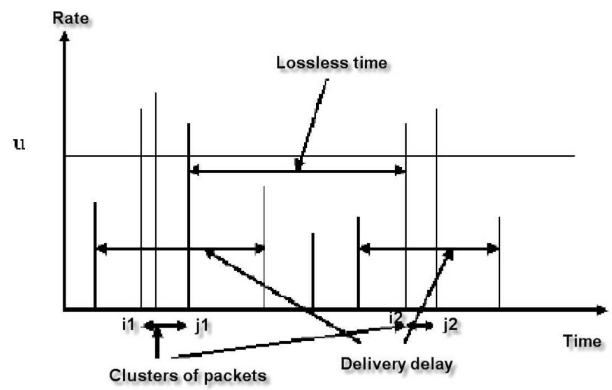
To develop our general methodology regarding the statistical analysis of dependent packet data, we subsequently consider a general packet flow and try to evaluate the quality of its transmission at a bottleneck link of the underlying packet-switched transport network. Let us assume that the corresponding IATs between the packets of a flow are random quantities X_i. We follow a fluid flow approach and describe the packet transport at the bottleneck link by a bufferless channel. We assume that a total packet loss is caused by the exceedances of the instantaneous transmission rates $R_i = Y_i/X_i$ of corresponding packets of lengths Y_i and IAT X_i over the available capacity u of the bufferless channel.

A bufferless system is a realistic model for real-time applications. Regarding the voice and video transmission the usage of a large packet buffer may lead to delays and a poor interpretation of the data at the receiver.[2]

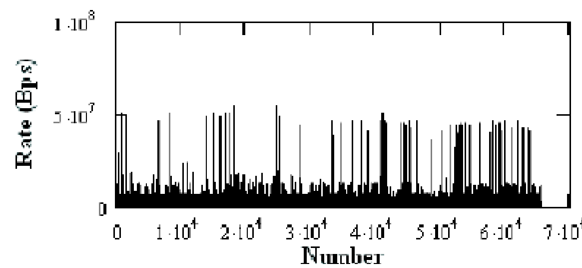
Such packets whose rates exceed the capacity threshold u occur in clusters (see Fig. 4(a)). Within the observation period we observe $N_2(u)$ such clusters that we enumerate by an index variable $C(u) \in \{1, \dots, N_2(u)\}$. We denote the first packet of a cluster $C(u)$ by its corresponding index $i_{C(u)}$ and its last packet by $j_{C(u)}$ (see Fig. 4(a)). We determine a cluster as a set of packets that are directly surrounded by two consecutive packets $i_{C(u)} - 1$ and $j_{C(u)} + 1$ whose corresponding rates do not exceed the capacity threshold u, whereas the rates of all packets $i_{C(u)}, \dots, j_{C(u)}$ in the cluster cross u. Fig. 4(b) depicts numerous bursts of the rates of an underlying P2P video stream during a SopCast session.

The underlying physical capacity u_p of the bottleneck link results from the settings of the transport network. The virtual capacity u is determined as an equivalent bandwidth of our recorded packet stream in the applied fluid flow approach. It may be equal to some quantile of the transmission rate of the packet stream. The logical transport channel can be used by different streams if $u < u_p$ holds.

[2]A channel with a buffer requires a special investigation and is beyond the scope of the current paper.



(a)



(b)

Figure 4: (a) Packet loss in terms of clusters of packets, the delay of a successful packet delivery due to loss in the clusters, and the lossless time corresponding to a selected channel capacity u. (b) Transmission rates arising from a video stream of a P2P SopCast session calculated by the ratios of the PLs to the IATs.

According to our fluid flow approach the delay of a successful packet delivery is determined by the time between the delivery of two consecutive packets whose related rates drop below the capacity u of the bufferless channel (see Fig. 4(a)). Formally, it is determined by a sum over a random number $T_1(u)$ of observations X_i of the generic r.v. X of the IATs,

$$d(u) = \sum_{i=1}^{T_1(u)} X_i \qquad (1)$$

Here the r.v.

$$T_1(u) = \min\{j \geq 1 : R_{j+1} \leq u | R_1 \leq u\} \qquad (2)$$

denotes the number of IATs between consecutive packets whose corresponding rates do not exceed the capacity threshold u. For simplicity, we re-enumerate here the IATs and PLs between non-exceedances of the process $\{R_i\}$ in (1) and (5), and further between exceedances in (3) and (4), respectively, such that they are starting from 1 again.

The delivery delay $d(u)$ may be caused by a packet loss in the clusters (for example, the time between R_1 and R_5 or R_7 and R_{10} in Fig. 4(a), where R_i denotes the rate of the ith packet) or by a packet gap due to a lack of peers with the required multimedia information or missing traffic from certain peers.

To exclude the latter delay, we consider the delay in a cluster $C(u)$

$$\begin{aligned} d^*(u) &= \sum_{i=1}^{T_{2,C(u)}(u)} X_i, \quad (3) \\ T_{2,C(u)}(u) &= \min\{j > 1 : R_{j+i_{C(u)}-1} \leq u | R_{i_{C(u)}-1} \leq u\} \end{aligned}$$

caused only by a packet loss in the clusters. Here, $T_{2,C(u)}(u) = j_{C(u)} - i_{C(u)} + 2$ is the random number of IATs between the packets in the cluster $C(u)$. Formula (3) differs from (1) since $T_{2,C(u)}(u)$ cannot be equal to 1.

The delay due to packet loss in the clusters highly impacts on the quality of experience that is perceived by a user observing the multimedia stream. The impact of a lack of packets and, hence, of the IATs on the perceived quality depends on the context. If traffic from certain peers is missing, it may cause severe difficulties for the receiver to decode the media frames, although the IATs could remain relatively small. However, the IATs and their quantiles could be even smaller if those packets were not lost. Hence, monitoring the IATs and their quantiles (see [10]) can be useful to control the quality of experience.

According to our fluid flow model of the channel the byte loss is only generated by clusters. It is determined by

$$b(u) = \sum_{i=1}^{T_{2,C(u)}(u)-1} Y_i, \quad (4)$$

where $T_{2,C(u)}(u) - 1 = j_{C(u)} - i_{C(u)} + 1$ denotes the number of packets in the cluster $C(u)$. This loss is caused by the packets $i_{C(u)}, \ldots, j_{C(u)}$ of lengths Y_i whose rates R_i exceed the capacity threshold u of the bufferless channel.

The lossless time (or the burstless period)

$$\begin{aligned} l(u) &= \sum_{i=1}^{T_{3,C(u)}(u)} X_i, \quad (5) \\ T_{3,C(u)}(u) &= \min\{i \geq 1 : R_{i+j_{C(u)}} > u | R_{j_{C(u)}} > u\} \end{aligned}$$

following a cluster $C(u)$ is the time between the arrivals of two consecutive packets $j_{C(u)}$ and $i_{C(u)+1}$ whose rates exceed the capacity threshold u. Here $T_3(u) = i = i_{C(u)+1} - j_{C(u)}$ coincides with the number of IATs between consecutive packets $j_{C(u)}$ and $i_{C(u)+1}$ whose corresponding rates exceed u.

The IATs within clusters determine also the lossless times. The IATs outside clusters may be equal to the delays $d(u)$.

During a monitored session one can get some number $N_2(u) \geq 1$ of clusters in the observations. Thus, we obtain corresponding sequences of delays

$$\begin{aligned} d_m(u) &= \sum_{i=1}^{T_{1,m}(u)} X_i, \quad m = \overline{1, N_1(u)}, \\ d_m^*(u) &= \sum_{i=1}^{T_{2,m}(u)} X_i, \quad m = \overline{1, N_2(u)}, \end{aligned}$$

byte losses

$$b_m(u) = \sum_{i=1}^{T_{2,m}(u)-1} Y_i, \quad m = \overline{1, N_2(u)},$$

and lossless (burstless) periods

$$l_m(u) = \sum_{i=1}^{T_{3,m}(u)} X_i, \quad m = \overline{1, N_3(u)}$$

arising within the time interval of the observation. Here N_1 denotes the number of delays, N_2 the number of clusters, and N_3 the number of burstless periods. The notation $T_{j,m}$, $j \in \{1, 2, 3\}$, denotes the corresponding values of $T_{j,\cdot}$ arising from the m-th delay, cluster and burstless period, respectively.

4.2 Analysis of Distributions, Quantiles and Means of the Quality Indices

Considering a fixed value of the capacity threshold u of a bottleneck link, the introduced performance characteristics $\{d_m(u)\}$, $\{d_m^*(u)\}$, $\{b_m(u)\}$, $\{l_m(u)\}$ related to the transmission of the considered packet flows to a selected peer are random variables (r.v.s). We assume in the following that the stationarity property has been validated with regard to these random entities (cf. [10]).

To study the arising performance isssues in terms of the defined QoS indices such as the delay-loss profile and the required equivalent bandwidth of a flow subject to QoS constraints, it is necessary to analyze the distributions and means of these r.v.s, and to estimate their quantiles. The problem is that these characteristics as well as the IATs $\{X_i, i \in \mathbb{N}_n\}$ and PLs $\{Y_i, i \in \mathbb{N}_n\}$ of a flow of sample size n may be dependent r.v.s and that the underlying generic r.v. X governing these IATs may have infinite variance.

4.2.1 Declustering the Data

The statistical analysis of dependent, stationary data like the defined quality indices $\{d_m(u)\}$, $\{d_m^*(u)\}$, $\{b_m(u)\}$, $\{l_m(u)\}$ requires the application of a declustering procedure. It means that the partitioning of the data of a P2PTV trace into independent blocks is necessary before a classical statistical analysis can be applied. Then one can deal with representatives of these data blocks similar to independent data and apply appropriate statistical estimation techniques.

The main feature of dependence is determined by the presence of clusters or conglomerates in the data. The more observations we have, the more visible are the clusters (see Fig. 4). Since data in the clusters contain approximately the same information, increasing the sample size n does not substantially improve the accuracy of an estimation process. If $\theta \in [0, 1]$ denotes the extremal index of the corresponding time series (cf. [2]), it implies that only the $n\theta$-part of the observations are effective. Thus, one can construct estimates by representatives of the clusters, e.g., maxima within clusters (see Fig. 5).

Let M_n and $\widetilde{M}_n$ denote the maxima of a gathered sequence of dependent r.v.s $\{X_1, ..., X_n\}$ and of a sequence of associated independent r.v.s $\{\widetilde{X}_1, ..., \widetilde{X}_n\}$ with the same distribution function (df)

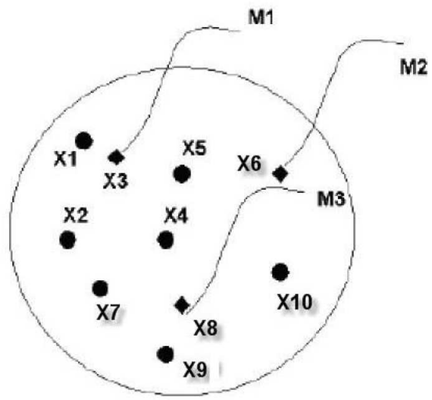


Figure 5: The independent block maxima $\{M_j, j = \overline{1, [n\theta]}\}$ of the initial data $\{X_i, i = \overline{1, n}\}$ consist of the $n\theta$-part of the observations and denote just some observations $\{X_i\}$ that are maximal within data blocks.

$F(x) = \mathbb{P}\{X \leq x\}$, respectively (cf. [2]). By the theory of extremes

$$\mathbb{P}\{M_n \leq u_n\} \approx \mathbb{P}^{\theta}\{\widetilde{M}_n \leq u_n\} = F^{n\theta}(u_n) \tag{6}$$

holds typically for sufficiently large sample size n and an appropriate level u_n.

Then (6) implies that one can operate with the dependent data by means of the extremal index θ in the same way like with independent ones. Regarding an i.i.d. sequence it holds $\theta = 1$. A value $\theta < 1$ provides some indication about the clustering behavior and, hence, a dependence in the underlying sequence. The reciprocal $1/\theta$ determines the mean size of the occuring clusters (cf. [2]).

To estimate the fundamental parameter θ, we can use the popular blocks and runs estimators (cf. [15]). Let

$$M_{i,j} = \max\{R_{i+1}, ..., R_j\}.$$

Let k denote the number of data blocks, and $r = [n/k]$ be the number of observations in each block where $[\cdot]$ denotes the integer part of a real number. Then the blocks estimator

$$\overline{\theta}^B(u) = \frac{n\sum_{j=1}^{k} \mathbf{1}(M_{(j-1)r,jr} > u)}{rk\sum_{i=1}^{n} \mathbf{1}(R_i > u)} \tag{7}$$

is based on the understanding that a cluster is a block of data with at least one exceedance over the threshold u (cf. [2], [15]). Here $\mathbf{1}(A)$ denotes the indicator function of the event A.

Regarding the runs estimator

$$\overline{\theta}^R(u) = \frac{(n-r)^{-1}\sum_{i=1}^{n-r} \mathbf{1}(X_i > u, M_{i+1,i+r} \leq u)}{n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i > u)}, \tag{8}$$

a cluster is defined as a block of data with some number of exceedances over the threshold u and at the same time the following r observations drop all below this threshold u. From a statistical perspective, the runs estimate is slightly superior since it has a better asymptotic bias than the blocks estimate (cf. [15]).

In practice, the optimal selections of the threshold u and the number of blocks k (or r) constitute a difficult issue that has no precise analytic answer up to now. It is the simplest way to estimate θ by that value $\widehat{\theta} \in \{\overline{\theta}^B(u), \overline{\theta}^R(u)\}$ which corresponds to the stability region of the plot $(u, \overline{\theta}(u))$ over a range of thresholds u.

4.2.2 Features of the Distributions

The distributions of $\{d_m(u)\}$, $\{d_m^*(u)\}$, $\{b_m(u)\}$, $\{l_m(u)\}$ could be Gaussian according to the central limit theorem if the number of terms in the sums increases and if the IATs $\{X_i\}$ and PLs $\{Y_i\}$ were independent and identically distributed (iid) r.v.s, each having finite variance. Evidently, the variance of the PLs is finite. The normality can follow if the PLs are weak dependent. This property is valid for the IPTV application SopCast (see Section 2, [10]) and hence, its byte loss is asymptotically Gaussian.

According to the generalized central limit theorem the distributions could be stable if the iid condition is fulfilled, but the variance of $\{X_i\}$ is infinite. The problem arises when $\{X_i\}$ and $\{Y_i\}$ are dependent, which is more realistic in practice. In this case the limit distribution of the sum may be stable under certain mixing conditions if the variance of the summands is infinite.[3] This implies that the variance of the sum will be infinite. The latter leads to the pessimistic conclusion that the delay may have asymptotically infinite variance (i.e., as the sample size increases to infinity) if the IATs are dependent with infinite variance. The optimistic conclusion is that the lossless time may also have an infinite variance. This is valid for the IPTV application SopCast whose IATs are found to be weak dependent with infinite variance (cf. [10]).

4.2.3 Quantile Estimation of the Quality Indices

Considering the continuous df $F(x) = \mathbb{P}\{D_1 \leq x\}$ of a r.v. D_1 the quantile $x = x_p$ of level $1-p$, $p \in (0, 1)$, is specified by definition as the solution of the equation $1-F(x_p) = \mathbb{P}\{D_1 > x_p\} = p$. To estimate the quantiles accurately, the observations of the corresponding r.v. have to be statistically independent. However, the underlying r.v.s of $\{d_m(u)\}$, $\{d_m^*(u)\}$, $\{b_m(u)\}$, $\{l_m(u)\}$ may be dependent for a sufficiently low threshold u. This is caused by the dependence of the clusters of packets that are located close to each other for low u.

To deal with independent observations, we have to perform a declustering of the data if it is possible. For this purpose, the data are separated into equal-sized blocks $J_1, ..., J_N$. Then one can find the maxima $M(J_j)$ within these blocks and estimate their quantiles. Namely, let

$$M_j = M(J_j) = \max\{D_{(j-1)s+1}, D_{(j-1)s+2}, ..., D_{js}\}, j = \overline{1, N}, \tag{9}$$

be the block maxima derived from the observations $\{D_m\}$ of some quality index. The size of each block s is a parameter that should be selected in such a way that the independence and a sufficient number of block maxima are provided. Unfortunately, checking the dependence cannot be done easily and automatically and requires several tests.

The dependence structure of the block maxima of each r.v. M_t modeling $\{d_m(u)\}$, $\{d_m^*(u)\}$, $\{b_m(u)\}$, $\{l_m(u)\}$ is derived from the sample autocorrelation function (ACF)

$$\widehat{\rho}(h) = \sum_{t=1}^{N-h}(M_t - \overline{M}_N)(M_{t+h} - \overline{M}_N) / \sum_{t=1}^{N}(M_t - \overline{M}_N)^2 \tag{10}$$

for the lag h, where $\overline{M}_N = (1/N)\sum_{i=1}^{N} M_i$ holds, the Ljung-Box's Portmanteau test (cf. [3]),

$$Q = N(N+2)\sum_{j=1}^{h} \hat{\rho}^2(j)/(N-j) \tag{11}$$

[3] See [1] for a survey on recent results regarding limit distributions of sums of dependent r.v.s..

and the Runde's Portmanteau test with tail index α (cf. [13]),

$$Q_R = (N/\ln N)^{2/\alpha} \sum_{j=1}^{h} \hat{\rho}^2(j), \qquad (12)$$

$N = [n/s]$ denotes the number of block maxima selected from the sample of size n, and $[\cdot]$ is the integer part of a real number. A selection procedure for the block size is described in Section 5 regarding the illustrative SopCast data.

The finiteness of the variance of the r.v. M_1 and, more generally, the heaviness of the tail of its distribution is an important criterion for the further selection of a test on independence of the observations. It is well known that for heavy-tailed distributions not all moments are finite in contrast to light-tailed distributions. Runde's test is appropriate if the variance of the generic r.v. M_1 is infinite, i.e. the distribution is heavy tailed. If the distribution of the generic r.v. M_1 is regularly varying, i.e. the distribution function (df) is given by

$$G(x) = 1 - \ell(x)x^{-\alpha}, \quad \alpha = 1/\gamma \geq 0,$$

where $\ell(x)$ is a slowly varying function, this corresponds to $0 < \alpha < 2$. The tail index α further shows the shape of the tail. It is known that α may indicate the number of finite moments of the r.v. M_1 if the distribution is regularly varying. Namely, $\mathbb{E}M_1^\beta < \infty$ is satisfied if $\beta < \alpha$ holds. The Pareto distribution $G(x) = 1 - (x/k)^{-\alpha}, x \geq k > 0$, for instance, is regularly varying.

The presence of regular variation and heavy (or light) tails of the r.v. M_1 can be checked by the sample mean excess function

$$e_N(u) = \sum_{i=1}^{N}(M_i - u)\mathbf{1}(M_i > u) / \sum_{i=1}^{N} \mathbf{1}(M_i > u). \qquad (13)$$

An increase of $e_N(u)$ implies a heavy-tailed distribution. Its linear increase indicates a Pareto-like regularly varying distribution. A constant value means that the distribution is exponential. The decrease implies that the distribution is light tailed.

Considering the specified quality indices, the estimation of the extremal index θ allows us to calculate their quantiles and the quantiles of their maxima. By (6) we realize that the $(1-p)$th quantile x_p of the maximum M_n corresponds to the $(1-p)^{1/(n\theta)}$th quantile of the distribution $F(x)$, i.e.

$$1 - p = \mathbb{P}\{M_n \leq x_p\} \approx F^{n\theta}(x_p).$$

We shall estimate high quantiles close to 100% to determine approximate upper bounds of the delivery delay, lossless time and byte loss.

We can suppose that the size s of the blocks is selected in such a way that the corresponding block maxima are independent. Then the quantiles of the quality indices D_i may be estimated by means of the quantiles of the independent block maxima M_j of these indices and we get:

$$1 - p = \mathbb{P}\{M_j \leq x_p\} \approx \mathbb{P}^{s\theta}\{D_1 \leq x_p\} = F^{s\theta}(x_p) \qquad (14)$$

It means that the $(1-p)^{s\theta}$th quantile x_p of the block maxima M_j corresponds to the $(1-p)$th quantile of a quality index D_m.

To estimate the $(1-p)^{s\theta}$th quantiles of the block maxima, we will use the well-known Weissman's estimator [18]

$$\widehat{x_p} = M_{(N-k_0)} \left(\frac{k_0 + 1}{(N+1)(1-(1-p)^{s\theta})} \right)^{\hat{\gamma}}, \qquad (15)$$

$k_0 = \overline{1, N-1}$. Here $M_{(1)} \leq M_{(2)} \leq \ldots \leq M_{(N)}$ denotes the order statistics of the sample $\{M_1, ..., M_N\}$ of the block maxima and $\hat{\gamma}$ is some estimate of their extreme value index γ. The latter is the reciprocal of the tail index α determining the shape of the df of the underlying r.v. M_j. The estimate (15) uses a Generalized Pareto distribution

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

$\sigma > 0, x \geq 0$ for $\gamma \geq 0$ and $0 \leq x \leq -\sigma/\gamma$ for $\gamma < 0$, as model of the tail of the corresponding distribution.

Regarding (15), only the k_0 largest statistics are used instead of the whole sample. The same k_0 can be used in the estimator $\hat{\gamma}$ of the block maxima M_j. We shall apply the popular Hill's estimator

$$\widehat{\gamma}^H(N, k_0) = \frac{1}{k_0} \sum_{i=1}^{k_0} \ln M_{(N-i+1)} - \ln M_{(N-k_0)} \qquad (16)$$

to estimate γ (cf. [7, p. 6]). One can select k_0 corresponding to the stability interval of the Hill's plot $\{k_0, \widehat{\gamma}^H(N, k_0)\}$, $k_0 = 1, 2, ..., N-1$ or automatically by a bootstrap method (see [7, Sec. 1.2]). The simplest way is to take $k_0 = [\sqrt{N}]$. The bootstrap estimate is obtained by averaging the Hill's estimates which are constructed over some number B of bootstrap re-samples from the underlying sample with repetitions. This minimizes approximately the mean squared error.

In conclusion, a generic algorithm regarding the quantile estimation of the quality indices looks as follows:

1. We determine a threshold u as some empirical quantile of the instantaneous transmission rates $R_i = Y_i/X_i$ using the sequences of the IATs $\{X_i\}$ and PLs $\{Y_i\}$ of a packet flow.
2. Based on the IATs and PLs of the packet traffic and the selected threshold u we construct samples of the general delivery delays $\{d_m(u)\}$ and delays within the clusters $\{d_m^*(u)\}$, the byte losses in the clusters $\{b_m(u)\}$ and the lossless times $\{l_m(u)\}$ by (1), (3), (4), (5).
3. We form data blocks of a selected size s such that we get statistically independent block maxima, see (9).
4. We estimate the extremal index θ of each sample by (7) and (8).
5. Finally, we estimate the quantiles of the corresponding quality indices by (14)-(16) using the quantiles of their block maxima M_j .

4.2.4 Mean Value Analysis

To estimate the means of the quality indices, we suppose that the conditions of Wald's equation are satisfied. Therefore, we suppose that the pairs of the r.v.s $(T_{j,m}(u), X_i)$, $j \in \{1, 2, 3\}$, and $(T_{2,m}(u), Y_i)$ are mutually independent, and $\mathbb{E}X_i < \infty$, $\mathbb{E}T_j(u) < \infty$, $j \in \{1, 2, 3\}$ hold.

Then the means of the delivery delay, the lossless time and the byte loss can be determined by Wald's equation which yields

$$\begin{aligned} \mathbb{E}(d_m(u)) &= \mathbb{E}(T_{1,m}(u))\mathbb{E}(X_i), \\ \mathbb{E}(d_m^*(u)) &= \mathbb{E}(T_{2,m}(u))\mathbb{E}(X_i), \\ \mathbb{E}(l_m(u)) &= \mathbb{E}(T_{3,m}(u))\mathbb{E}(X_i), \\ \mathbb{E}(b_m(u)) &= (\mathbb{E}(T_{2,m}(u)) - 1)\mathbb{E}(Y_i). \end{aligned}$$

The corresponding sample estimates can be calculated by

$$\bar{d}(u) = \overline{T_{1,m}}(u)\overline{X}, \qquad (17)$$

$$\overline{d^*}(u) = \overline{T_{2,m}}(u)\overline{X^*} \qquad (18)$$

$$\bar{l}(u) = \overline{T_{3,m}}(u)\overline{X}, \qquad (19)$$

$$\bar{b}(u) = (\overline{T_{2,m}}(u) - 1)\overline{Y^*}, \qquad (20)$$

where $\overline{X}$, $\overline{T_{j,m}}(u)$, $j \in \{1,2,3\}$ are the sample averages of the IATs $\{X_i\}$ and $\{T_{j,m}(u)\}$, respectively. $\overline{X^*}$ and $\overline{Y^*}$ are taken over the IATs and PLs within clusters. $\overline{X^*} \neq \overline{X}$ holds since the IATs within clusters may be on average different from the IATs corresponding to all packets. For example, $\overline{X^*} < \overline{X}$ holds regarding our SopCast data since video packets that form clusters are more frequent than signaling packets. Similarly, $\overline{Y^*} > \overline{Y}$ follows since clusters contain video packets that are larger than the signaling packets.

To use the formulas (17)-(20), we have to check the pairwise independence between the r.v.s $T_j(u)$ and X_i, $j \in \{1,2,3\}$ and, $T_2(u)$ and Y_i, and the finiteness of $\mathbb{E}X_i$, $\mathbb{E}T_j$. Since $T_j(u), j \in \{1,2,3\}$, depend on the rate, one has to check the dependence between the IATs X_i and the transmission rates R_i and between the PLs Y_i and the rates. This can be done by an estimation of copulas and Pickand's function (see [7, 8, 11]) and it is beyond the scope of the paper.

The independence of the duration of a connection and the transmission rate is shown to be realistic for streaming media and P2P networks since the transmission durations are given by the lifetime of a user on the P2P networks, while the rates are given by the maximum upload bandwidth (cf. [4]). In [8] the approximate mutual independence of the IATs and the rates and the dependence between the PLs and rates have been revealed for the SopCast packet flow investigated here. Hence, formula (20) cannot be used for the considered SopCast data. However, one can use (20) for any other IPTV applications when their PLs and rates are mutually independent.

The sample average is a reliable estimate of the mean if the observations are independent or weak dependent and the expected value of the underlying r.v. is finite. For example, we can estimate $\mathbb{E}(X_i)$ and $\mathbb{E}(Y_i)$ by $\overline{X}$ and $\overline{Y}$ in case of a SopCast flow since the first moments of the IATs and PLs are shown to be finite and the IATs and PLs are weak dependent r.v.s (see [10]). $\overline{X}$ and $\overline{Y}$ cannot be used if the IATs and PLs are heavy long-range dependent sequences. To overcome this problem, one should decluster the data (see subsection 4.2.1) and estimate the mean of the block maxima as an upper bound of the expected value. It is worthwhile to mention that the distributions of $T_j(u)$ are asymptotically close to a geometric distribution for sufficiently high thresholds u (see [9]).

5. APPLICATION OF THE METHODOLOGY TO REAL P2PTV TRAFFIC

In the following we shall apply the proposed statistical methodology to the packet traffic gathered at an observed single peer in the P2P overlay network during a Sopcast session.

5.1 Data Analysis of a P2P Traffic Flow

We have applied the declustering approach and separated the samples $\{d_m(u)\}$, $\{d^*_m(u)\}$, $\{b_m(u)\}$ and $\{l_m(u)\}$ into blocks of equal size. The corresponding block sizes (i.e., the number of observations in the block) are indicated in Table 2. Regarding a relatively low capacity level u independent block maxima cannot be obtained due to the strong dependence between the data blocks in this case. Therefore, we have selected $u = 10^7$ bps which corresponds to the 99.2% quantile of the instantaneous flow rate $R_i = Y_i/X_i$ for our SopCast data (see Fig. 4(b)). The descriptive statistics of $\{d_m(u)\}$, $\{d^*_m(u)\}$, $\{b_m(u)\}$ and $\{l_m(u)\}$ for this u are summarized in Table 2.

Further we need to check the independence of the corresponding block maxima. To select the right independence test, one has to evaluate first the number of finite moments of the block maxima. This can be done by means of the tail index and the mean excess function (13). The tail index $\alpha = 1/\gamma$ of the corresponding block maxima M_j has been calculated by a combination of the Hill's estimator (16) and the bootstrap method with $B = 100$, see Table 2 (cf. [7]). The positive sign of the tail index implies that the tails of the distributions of all considered quality indices are heavy.

We have also checked the mean excess functions (13) of the block maxima of $\{d_m(u)\}$, $\{d^*_m(u)\}$, $\{b_m(u)\}$ and $\{l_m(u)\}$. Assuming regular variation, one may conclude from Fig. 6 that the distribution of the block maxima of $\{d^*_m(u)\}$ is Pareto-like, that of $\{b_m(u)\}$ is a mixture of light-tailed and some heavy-tailed distributions (the latter is in the agreement with the asymptotic normality predicted in Section 4.2.2) and those of $\{d_m(u)\}$ and $\{l_m(u)\}$ are mixtures of heavy- and light-tailed components with a domination of the heavy-tailed regularly varying ones.

Regarding Table 2 one can state that the distribution of the block maxima of the delivery delay in the clusters $\{d^*_m(u)\}$ is heavy tailed and has infinite variance and mean since $\alpha = 0.77 < 1$ holds. The block maxima of all other indices are heavy tailed with finite variances since their tail indices satisfy $\alpha > 2$.

The "extraction" of the delay in clusters $\{d^*_m(u)\}$ allows us to make the pessimistic conclusion that the mean of the maximum of the delivery delay due to the loss of packets in the clusters d^*_m may be infinite for the selected threshold u.

The obtained block maxima have also been investigated regarding their independence using the ACF (see Fig. 7), and the Ljung-Box test as well as Runde's test (see Tab. 3). The sample ACF (10) of the block maxima of $\{d^*_m(u)\}$ cannot be calculated since the mean and variance of these block maxima are infinite. Therefore, we have used the modified estimate without a centering by the sample mean (see [12]):

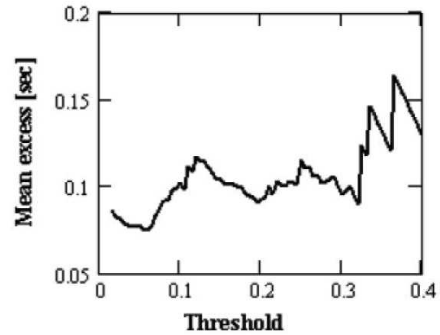
$$\widehat{\rho}(h) = \sum_{t=1}^{N-h} M_t M_{t+h} / \sum_{t=1}^{N} (M_t)^2.$$

Since the ACFs are all located inside the Gaussian confidence interval (apart of a few points of the ACF regarding the block maxima of $\{d_m(u)\}$), one can expect that the block maxima may be independent. This independence is derived by the Portmanteau tests (11) and (12). Since the variance of the block maxima of $\{d^*_m(u)\}$ is infinite, one has to check the independence of the block maxima by Runde's test. The block maxima of all other indices are checked by the Ljung-Box test. Since the values of Runde's statistic Q_R do not exceed the critical values $Q_h(0.05)$ of the limit distribution of Q_R for the 0.05 level given in [13], the null hypothesis regarding the independence of the block maxima of $\{d^*_m(u)\}$ should be accepted. Since the values Q do not exceed $\chi^2_{0.05}(h)$, where $\chi^2_\eta(h)$ denotes the ηth quantile of the chi-square distribution with h degrees of freedom, i.e., $Pr\{\chi^2 > \chi^2_\eta(h)\} = \eta$, $0 < \eta < 1$, the null hypothesis regarding the independence of the block maxima of $\{d_m(u)\}$, $\{b_m(u)\}$ and $\{l_m(u)\}$ should be accepted, too, see Table 3.

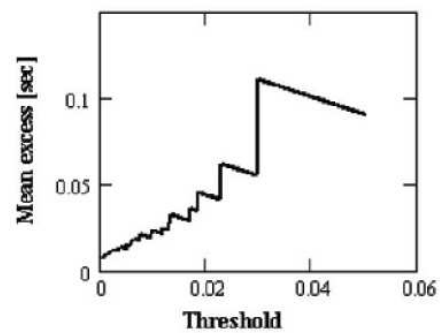
In this manner we have proved the independence of the block maxima of the quality indices and can estimate the high quantiles (99% and 99.9%) of these indices by formulas (14)-(16). For this purpose we have additionally estimated the extremal index θ of each quality characteristic by the estimators (7) and (8), where $k = r = [\sqrt{n}]$ is taken for the sample size n of a quality index, see Fig. 8. The corresponding results are stated in Table 2. Note that the quantiles may be larger than the maximal values of the quality indices.

Table 2: Statistics of $\{d_m(u)\}$, $\{d^*_m(u)\}$, $\{l_m(u)\}$ [sec] and $\{b_m(u)\}$ [kB] of the P2P packet traffic and their block maxima given the threshold $u = 10^7$ bps.

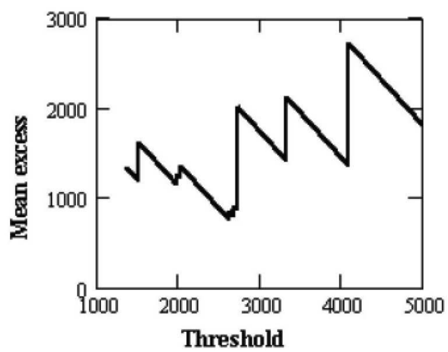
R.V.	Sample size	Min	Max	Median	Mean	Block size s	Tail index α of block maxima	Variance of block maxima	θ : $\overline{\theta}^B$/ $\overline{\theta}^R$	Quantiles 99%	Quantiles 99.9%
$d_m(u)$	65010	$2.1 \cdot 10^{-5}$	0.625	$5.68 \cdot 10^{-4}$	$4.974 \cdot 10^{-3}$	200	3.09	finite	0.45	0.358	0.628
$d^*_m(u)$	187	$5.66 \cdot 10^{-4}$	0.141	$3.43 \cdot 10^{-4}$	$2.55 \cdot 10^{-3}$	4	0.77	infinite	0.51	0.185	3.654
$b_m(u)$	187	0.640	6.81	1.092	0.801	4	15.132	finite	0.65/	1.668	1.941
									0.29	1.758	2.047
$l_m(u)$	524	$2.6 \cdot 10^{-5}$	14.254	0.017	0.613	7	2.149	finite	0.58	6.528	19.933



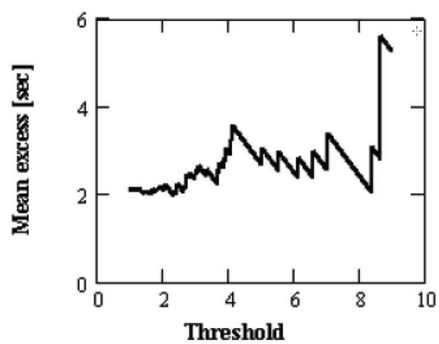
(a)



(b)



(c)



(d)

Figure 6: Sample mean excess functions of the block maxima of (a) $\{d_m(u)\}$, (b) $\{d^*_m(u)\}$, (c) $\{b_m(u)\}$ and (d) $\{l_m(u)\}$ of the composite SopCast traffic.

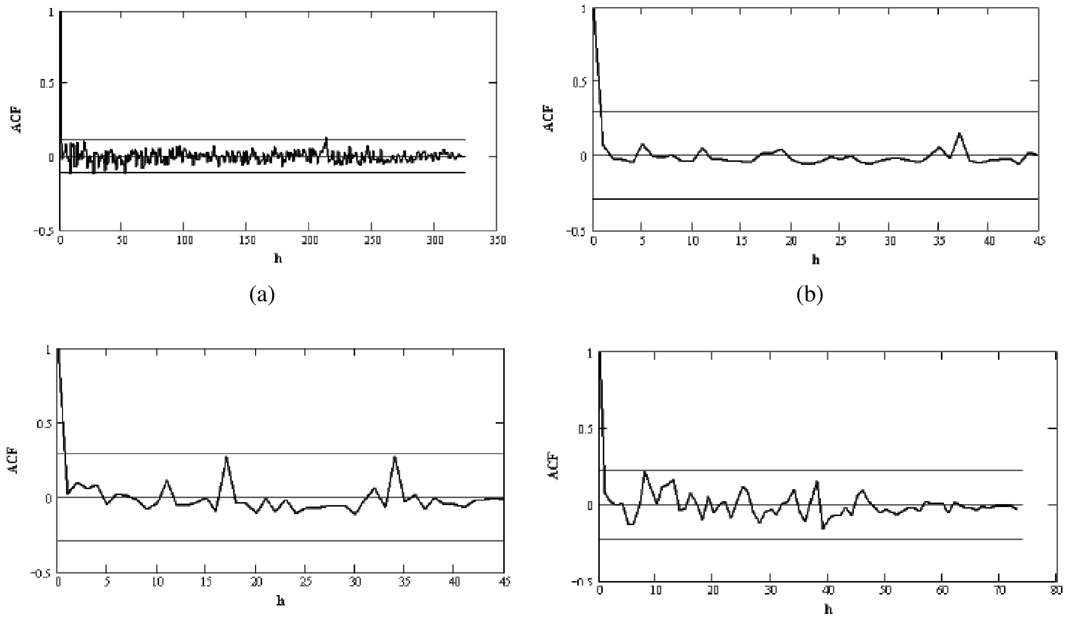
5.2 Mean Delay of Successful Packet Delivery and Mean Lossless Time

We have calculated the mean delay of the packet delivery by (17) and the mean lossless time by (19), both for the composite P2PTV traffic and the pure video traffic. It is impossible to calculate the mean byte loss per cluster by (20) due to the dependence of the PLs and the rates derived for the SopCast data in [8]. The estimate $\overline{d^*}(u)$ indicates the mean delay arising in the clusters of packets due to the exceedances of the rates over the available capacity u. The estimate $\overline{d}(u)$ indicates the mean delay caused by different reasons including the loss in the clusters, the missing of packets and silence periods.

The mean delay (see Fig. 9(a)) tends to decrease (the delay in the clusters does not decrease in a monotone fashion) and the mean lossless time increases (see Fig. 9(b)) as the quantile of the rate u increases, see Table 4. Here u may be considered as the equivalent bandwith that is required by the considered P2PTV flow on a bufferless channel subject to QoS constraints on loss and delay. The video traffic obeys a higher mean delay than the composite traffic, see Fig. 9(a). This behavior was predictable since the signaling traffic with small PLs and large IATs generates small rates which mostly cannot exceed the assigned capacity threshold and, thus, not generate any clusters of rate exceedances.

The mean lossless time of pure video traffic is larger than that of the composite traffic, see Fig. 9(b) and Table 4. Their ratio increases up to $u = 32 \cdot 10^5$ bps and stabilizes for $u > 10^7$ bps, see Fig. 9(c).

The mean delay $\overline{d}(u)$ of the composite and video traffic tends to the mean IAT for sufficiently large capacity (for u larger than or equal to the 99%th quantile of the rate, see Table 1) since $\overline{T_{1,m}}(u) \approx 1$ holds both for pure video and the composite traffic. The latter implies the existence of a few clusters with a few packets for a high threshold u. The mean delay in the clusters $\overline{d^*}(u)$ of both the composite and pure video traffic tends to zero (slower for video traffic)



(a) (b) (c) (d)

Figure 7: ACFs of the block maxima of (a) $\{d_m(u)\}$, (b) $\{d^*_m(u)\}$, (c) $\{b_m(u)\}$ and (d) $\{l_m(u)\}$ of the composite SopCast traffic given a Gaussian 95% confidence interval with the bounds $\pm 1.96/\sqrt{n}$ and the sample size n of the corresponding quality characteristic.

Table 3: Results on block maxima by the Ljung-Box test and Runde's test.

Lags	**Q of Block maxima**			$\chi^2_{0.05}$	**Lags**	$\mathbf{Q_R}$ **of Block**	$\mathbf{Q_h}$
	$d_m(u)$	$b_m(u)$	$l_m(u)$	(h)		**maxima** $d_m^{ast}(u)$	(0.05)
10	14.722	1.761	8.242	18.3	2	5.518	13.53
20	29.362	10.339	15.827	31.4	3	6.228	16.32
30	38.882	15.956	21.93	43.8	4	8.066	18.28
					5	13.708	19.17

Table 4: The mean lossless time $\bar{l}$ and, the mean delivery delays $\bar{d}$ and $\overline{d^*}$ for the total and the pure video SopCast traffic and a channel capacity u determined by the empirical quantiles of the rate.

	Quality index	**Rate quantiles**				
		75%	95%	97%	99%	99.6%
Total traffic	u (kbps)	1370	5260	5680	7880	14000
	$\bar{l}$ (msec)	20	99	167	471	1206
	$\bar{d}$ (msec)	6.575	5.193	5.084	4.986	4.954
	$\overline{d^*}$ (msec)	5.153	3.566	3.784	5.651	0.94
Video traffic	u (kbps)	733.4	2180	2900	14000	43935
	$\bar{l}$ (msec)	72	362	605	1694	3638
	$\bar{d}$ (msec)	24	19	19	18	18
	$\overline{d^*}$ (msec)	29	30	25	17	16

due to the disappearance of the clusters. The mean delay is larger for the video traffic since its mean IAT is larger after the virtual exclusion of the signaling packets. The difference between $\bar{d}(u)$ and $\overline{d^*}(u)$ is more significant for the composite traffic. For video traffic the estimates are similar since video packets are located mostly in the clusters. Without the signaling traffic the mean delay of the successful packet delivery and the mean lossless time would be larger than in the presence of the signaling traffic.

6. CONCLUSIONS

In our study measurements that are arising from a next generation network have been considered. We have presented advanced mathematical techniques to cope with the statistical analysis and modeling of these gathered data. This improvement is required to study important performance indices of new real-time services in a multimedia Internet such as the demanded bandwidth or delay-loss profiles of packet flows arising from a peer-to-peer overlay network during a multimedia session. The latter data sets incorporate strongly correlated or long-range dependent time series and heavy-tailed marginal distributions determining the underlying random variables of the data features such as the inter-arrival times or instantaneous transmission rates of the packets.

We have used a fluid flow approach based on the model of a bufferless transmission channel to characterize the bandwidth demand of a source flow with variable bitrate. The exceedances of the instantaneous transmission rates of a flow over the capacity threshold which are determined by the ratios of the inter-arrival times and associated packet lengths of a flow occur in terms of clusters. They are considered as the main source of the packet loss and the delay of a successful packet delivery.

Following these ideas, we have defined new quality indices with regard to the packet transmission of a multimedia flow. The latter are specified in terms of the distributions, high quantiles and the means of the delivery delay between successfully transmitted pack-

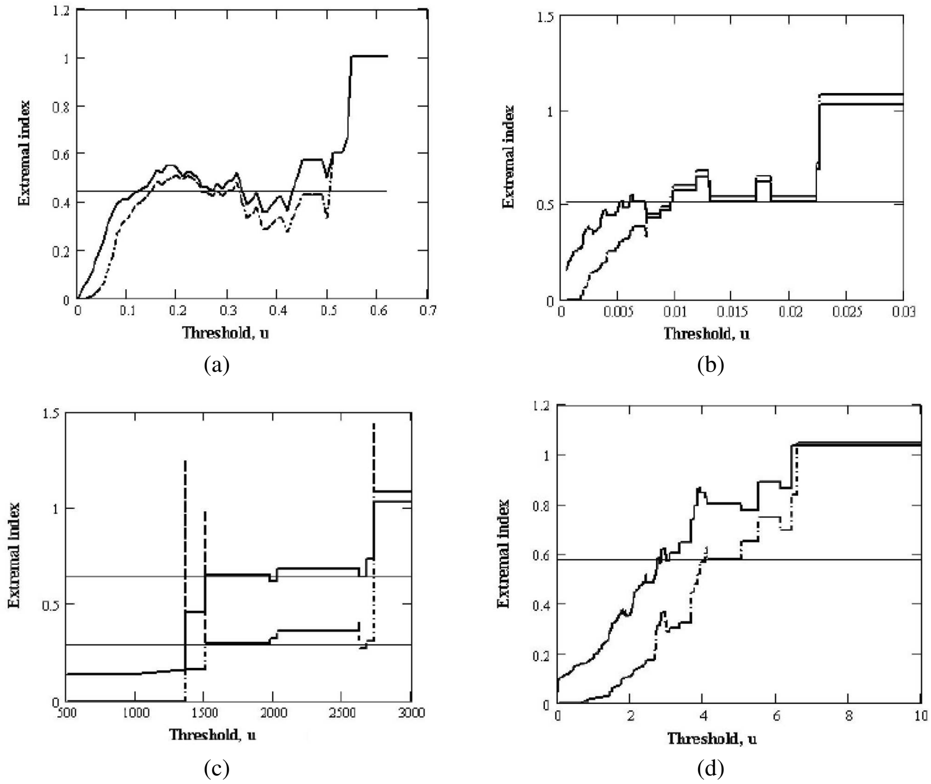


Figure 8: The blocks (solid line) and runs (dashed line) estimates (7) and (8) of the extremal index of (a) $\{d_m(u)\}$, (b) $\{d^*_m(u)\}$, (c) $\{b_m(u)\}$, and (d) $\{l_m(u)\}$ against the threshold u. The intervals of approximate stability correspond to the estimate of the extremal index (thin solid line).

ets, the lossless time and the byte loss. The clusters of threshold exceedances of the packet proceess are caused by the statistical dependence of the packets in an aggregated flow transmitted towards a destination. Hence, the impact of the dependence of the flow data and the heavy-tailed distributions of the underlying basic random variables on the QoS indices can be evaluated by our methodology. The advantage of the proposed methodology is given by its applicability to dependent data with infinite variance and mean.

To illustrate the new methodology, we have investigated the packet flows that are exchanged with a single peer and arising from the P2PTV application SopCast. The high quantiles of the defined quality indices are estimated for the composite packet stream including signaling packets and video content. Further the means of the delivery delay and the lossless time are compared both for the composite traffic and the unadulterated video traffic. Regarding the SopCast application the following conclusions can be drawn:

- Increasing the required equivalent bandwith of a flow leads to a decreasing mean delivery delay and to an increasing mean lossless time. This mean delay caused by different reasons tends to the mean inter-arrival time. The mean delay caused by the packet loss in the clusters due to the exceedance of the bandwith threshold tends to zero due to the disappearance of these clusters as expected.
- The quantiles of the delay show that the largest delivery delay is caused by the clusters of those packets whose rates exceed the equivalent capacity of the flow.
- The delivery delay and the lossless time may have an asymptotically infinite variance as the sample size increases to infinity due to the dependence and heavy tail of the packet inter-arrival process.
- The byte loss is asymptotically Gaussian as the sample size increases since the packet lengths are weak dependent and their variance is finite.
- If the signaling traffic were not considered in the analysis, it would increase the mean delivery delay and the mean lossless time. A larger mean delay due to a lack of signaling packets does not make the visual interpretation worse.

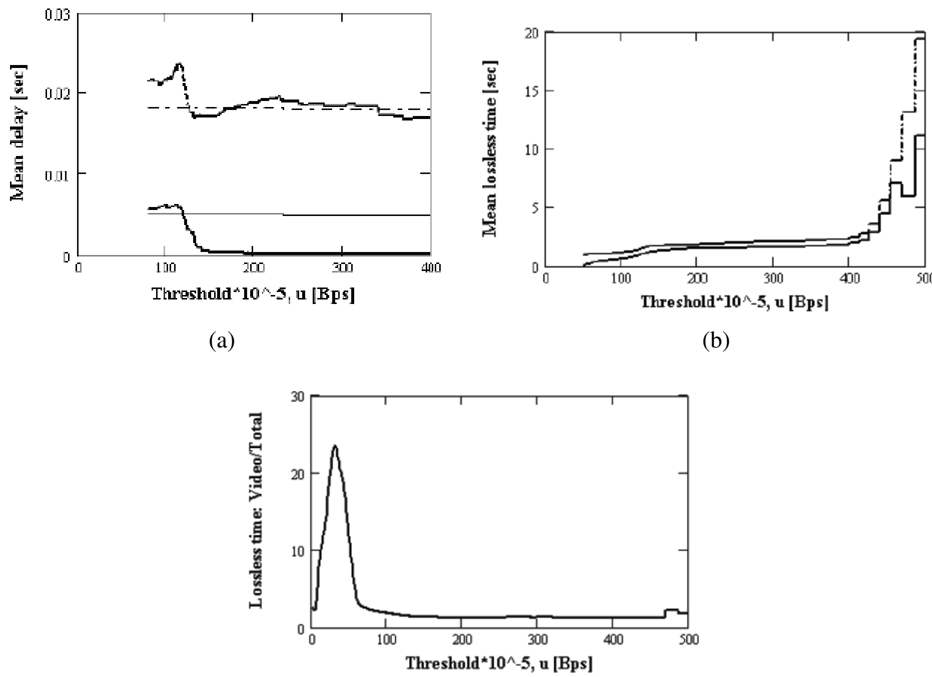
We are convinced that the presented concept can help to provide a mathematically well defined methodology to analyze measurements of distributed systems and, in particular, to evaluate relevant QoS indices of new multimedia applications with variable bitrate exclusively by means of these measured data. We have further shown that this universal concept can also be applied to data with underlying dependences and heavy-tailed distributions.

Acknowledgment

The authors acknowledge the partial support by a research grant of COST Action IC0703 "Data Traffic Monitoring and Analysis (TMA)". They appreciate the efforts of Mr. Schwessinger and Mr. Eittenberger at Otto-Friedrich University Bamberg during the P2P traffic measurement campaigns.

7. REFERENCES

[1] B. Basrak, D. Krizmanić, and J. Segers. A functional limit theorem for partial sums of dependent random variables with



(c)

Figure 9: The mean delay per cluster of the composite and pure video traffic: (a) a comparison of $\overline{d}$ (thin lines) and $\overline{d^*}$ (thick lines) calculated by (17); (b) the mean lossless time $\overline{l}(u)$ calculated by (19). In all figures, the composite traffic is marked by a solid line and the unadulterated video traffic by a dashed line. (c) Ratio of the mean lossless time of the video traffic to that one of the total traffic.

infinite variance. DP1001. arXiv: 1001.1345v1 [math.PR], 2010.

[2] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of Extremes: Theory and Applications*. J. Wiley & Sons, Chichester, 2004.

[3] P. J. Brockwell, R. A. Davis. *Introduction to Time Series and Forecasting*. Springer Texts in Statistics, New York, 2nd edition, 2002.

[4] B. D'Auria, S. I. Resnick. The influence of dependence on data network models of burstiness. *Advances in Applied Probability*, 40(1): 60–94, 2008.

[5] L. Ni, A. Harwood. A comparative study on peer-to-peer failure rate estimation. *Proc. International Conference on Parallel and Distributed Systems*, Vol. 2, pages 1–7, December 5–7, 2007.

[6] F. Liu, Z. Li. A Measurement and Modeling Study of P2P IPTV Applications. *Proceedings 2008 International Conference on Computational Intelligence and Security* 1, pages 114–119, 2008.

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

[8] N.M. Markovich. Modeling of Dependence in a Peer-to-Peer Video Application. *Proceedings International Wireless Communications and Mobile Computing Conference (IWCMC 2010)*, Caen, France, 2010.

[9] N.M. Markovich. Modeling of dependence by moderate samples. *Proceedings International Conference "Modern Stochastics: Theory and Applications II" (MSTA II 2010)*, Kyiv, Ukraine, September 7–11, 2010.

[10] N. M. Markovich, U. R. Krieger. Statistical Analysis and Modeling of Peer-to-Peer Multimedia Traffic. In D. Kouvatsos (ed.), *Next Generation Internet: Performance Evaluation and Applications*, LNCS 5233, Springer, Heidelberg, September 2010.

[11] R.B. Nelsen. *An Introduction to Copulas*. Springer, New York, 1998.

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

[13] R. Runde. The asymptotic null distribution of the Box-Pierce Q-statistic for random variables with infinite variance. *J. of Econometrics*, 78:205–216, 1997.

[14] T. Silverston, O. Fourmaux, A. Botta, A. Dainotti, A. Pescapé, G. Ventre, K. Salamatian. Traffic analysis of peer-to-peer IPTV communities. *Computer Networks*, 53:470–484, 2009.

[15] R.L. Smith, I. Weissman. Estimating the Extremal Index. *J. R. Statist. Soc. B.*, 56(3):515–528, 1994.

[16] S. Tang, Y. Lu, J.M. Hernández, F. Kuipers, P.Van Mieghem. Topology Dynamics in a P2PTV Networks. In L. Fratta et al. (eds.), *Networking 2009*, Aachen, 11-15 May 2009, pages 326–337, LNCS 5550, Springer, 2009.

[17] V. Zaborovsky, R. Meylanov. Peer-to peer fractal models: new approach to describe multiscale network processes, *Proceedings ICT 2002*, Beijing, 2002.

[18] I. Weissman. Estimation of parameters and large quantiles based on the k largest observations. *Journal of American Statistical Association*, 73:812–815, 1978.