Statistical Analysis of the End-to-End Delay of Packet Transfers in a Peer-to-Peer Network

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

[1] V.A. Trapeznikov Institute of Control Sciences Russian Academy of Sciences, Profsoyuznaya Street 65, 117997 Moscow, Russia
markovic@ipu.rssi.ru,nat.markovich@gmail.com
[2] Fakultät WIAI, Otto-Friedrich-Universität, An der Weberei 5, 96047 Bamberg, Germany
udo.krieger@ieee.org

Abstract. The paper is devoted to the statistical analysis of the end-to-end (E2E) delay of packet transfers between source and destination nodes in a peer-to-peer (P2P) overlay network. We focus on the identification of the E2E delay and the longest per-hop delay distributions and the stochastic dependence of the associated random process. The E2E delay is determined by the sum of a random number of dependent per-hop (p-h) delays along the links of a considered overlay path and the longest per-hop delay by their maximum. We propose to use the sum of the p-h delays to get a distribution of the maximum which is motivated by the available statistical data of the E2E delays. Based on recent analytic results derived from extreme-value theory we show that such sums and maxima corresponding to different paths may have the same tail and extremal indexes. These indexes determine the heaviness of the distribution tail and the dependence of extremes. Using the extremal index we identify limit distributions of the maxima of the E2E delays and the maxima of the p-h delays at a path among all source-destination paths. Considering real-time applications with stringent E2E-delay constraints, the distributions are used to identify quality-of-service (QoS) metrics of a P2P model like the packet missing probability and the corresponding playback delay as well as the equivalent capacity of a transport channel.

Keywords: P2P network · End-to-end delay · Per-hop delay · Tail index · Extremal index · Quality-of-service · Packet missing probability · Playback delay · Equivalent capacity

1 Introduction

We consider the delay performance of the packet transfer in a peer-to-peer (P2P) overlay network. The identification of the distribution of the end-to-end (E2E) delays arising between source and destination nodes in a P2P network constitutes an important problem of telecommunication due to live TV and video-on-demand applications. The delay of information transmission through the P2P

V. M. Vishnevskiy et al. (Eds.): DCCN 2020, LNCS 12563, pp. 282–297, 2020.
https://doi.org/10.1007/978-3-030-66471-8_22

network and, hence, the playback delay that is the lag between the generation of a packet and its playout deadline have a big impact on the quality of service and experience. As the E2E delay can be represented as a sum of a random number of the per-hop (p-h) delays, its distribution depends on the distributions of the random length of the overlay path between the source and destination and the p-h delays. The latter are determined by the structure of the P2P overlay network.

In [5], [15] the relation between the distribution of the packet delay and the packet missing probability in a P2P network has been considered. The distribution of the E2E delay of the ith path $D_i(D) = \sum_{j=1}^{L_i(D)} X_{i,j}$ is required. Here, $\{X_{i,j} : 1 \le j \le L_i(D)\}$ are the p-h delays of this overlay path i from the source S to the destination node D with a random length $L_i(D)$. The paths between S and D are schematically shown in Fig. 1. Their randomness is caused by the random number of nodes and links of the paths due to the dynamics of the P2P network over the time. The exceedance of the packet delay over the playback deadline b is considered as one of the main reasons to miss a packet. Then this part of the missing probability is the following: $P_m(b) = P\{D_i(D) > b\}$. Considering the E2E delays, we deal here with the sums of a random number of terms which can be heavy-tailed distributed and dependent. These issues constitute a complicated mathematical problem. In [15] the exceedance of the realized packet transmission rate over the equivalent capacity of the transport channel is considered as the second reason to loose packets.

It is one of the objectives of our paper to identify the missing probability under more general assumptions than in [5], [15] in view of the last statistical results obtained in [18]. It was assumed in [5], [15] as well as in [26] that $\{X_{i,j}\}$ are independent and identically distributed (i.i.d.) random variables (r.v.s) with light or heavy tails depending on the P2P overlay structure, and that the number of nodes N in the network and $L_i(D)$ are stationary distributed. The mutual dependence or independence of $X_{i,j}$ and $L_i(D)$ and the assumption which tail of these r.v.s is heavier are essential in order to identify the distribution of the sum, see for instance [8]. Here we assume that the p-h delays $\{X_{i,j}\}$ are now not necessarily i.i.d.. This assumption is realistic since paths may be overlapping as in Fig. 1. We assume that $\{X_{i,j}\}$ are stationary distributed at links located at the same distance with regard to the number of links from S. The random path length $L_i(D)$ is assumed to be stationary distributed, but its mutual independence on the p-h delay is omitted.

Another objective is to find the relation between the local dependence (i.e. cluster) structure and the distributions of the E2E delay and the maximal p-h delay at a path. This allows us to generalize the probability $P_m(b)$ uniformly to all paths of lengths $\{L_i\}$ and to obtain $P\{\max_i D_i(D) > b\}$. Our achievements are based on the results of extreme-value theory obtained in [18].

In [18] it is derived that the tail index (TI) and extremal index (EI) of the asymptotic distributions of sums and maxima of random length sequences are the same subject to some not very restrictive assumptions. One may conclude that the sums and maxima of p-h delays at the paths have the same heaviness

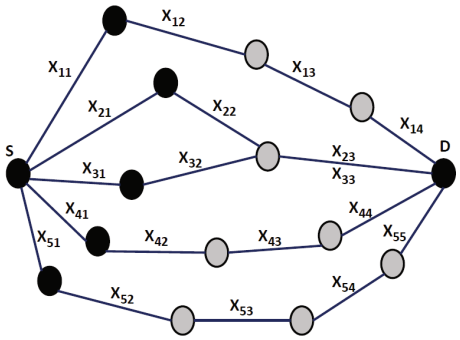

Fig. 1. Paths of random length between source node S and destination node D with the per-hop time delays $\{X_{i,j}\}$ of packet transmissions on the ith path between these nodes; the nodes in black between S and D indicate those ones with a distance of one link from S.

of the distribution tail and the same dependence structure. This feature implies that the distribution of the E2E delay may be approximated asymptotically by the distribution of the maximum of the p-h delays at a source-destination path. As the E2E delays can be made available in practice easier than the p-h delays, this allows us to approximate the distribution of the p-h delays at the most heavy-tailed link using the E2E delay statistics. Then the common TI value and the common EI value can be estimated by a sample of the observed E2E delays. The EI allows us to obtain the common limiting distribution of both the maxima of the E2E delays and the longest p-h delays among all paths. Moreover, one can use the distribution of the maximum to determine the packet missing probability.

The paper is organized as follows. Section 2 contains a survey of related results. In Sect. 3 our main results related to the stochastic model, its nonparametric estimation using a basic statistical algorithm, as well as an illustrative computational example are presented. The exposition is finalized with some conclusions and the discussion of open problems.

2 Related Work

Let the links of a path in the P2P network be enumerated from the source node S (see Fig. 1). We assume that the p-h delay $X_{i,j}$, $i, j \geq 1$, at the link j of path i is regularly varying distributed in a uniform way. This assumption implies that

$$\mathbb{P}\{X_{i,j} > x\} = \ell_j(x)x^{-k_j} \tag{1}$$

holds with the TI k_j and a slowly varying function $\ell_j(x)$, i.e. $\lim_{x\to\infty} \ell_j(tx)/\ell_j(x) = 1$ for any $t > 0$. Positive constants and logarithms provide examples of slowly varying functions $\ell_j(x)$. The links with the same number j are assumed

to be stationary distributed and their distributions may be different from the distribution of the links with another number.

The EI θ is sometimes called the local dependence measure having in mind that extremes or consecutive exceedances over a high threshold u occur usually in clusters. Such clusters of exceedances are caused by the dependence in stochastic sequences. The clustering can be intensified by heavy distribution tails.

Definition 1. *[13] The stationary sequence of r.v.s* $\{X_n\}_{n\geq 1}$ *with cumulative distribution function (cdf)* $F(x)$ *and* $M_n = \max\{X_1, ..., X_n\}$ *is said to have the EI* $\theta \in [0, 1]$ *if for each* $0 < \tau < \infty$ *there is a sequence of real numbers* $u_n = u_n(\tau)$ *such that it holds*

$$\lim_{n\to\infty} n(1 - F(u_n)) = \tau, \qquad \lim_{n\to\infty} P\{M_n \leq u_n\} = e^{-\tau\theta}. \tag{2}$$

The inverse $1/\theta$ approximates asymptotically the mean cluster size, i.e. the mean number of exceedances per cluster [13]. The cluster structure of a simulated Moving Maxima process [1], for instance, is shown in Fig. 2. The details regarding this process are recalled in Sect. 3.3. A smaller θ corresponds to wider clusters. In this example the values $\theta = 0.3$ and $\theta = 0.8$ imply that the mean cluster may contain approximately 3 and 1 exceedances, respectively.

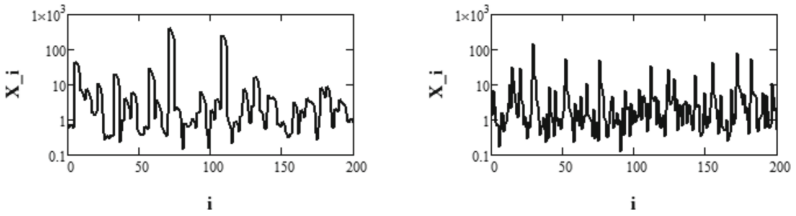

Fig. 2. The Moving Maxima process with larger and smaller clusters of exceedances for the EIs $\theta = 0.3$ (left) and $\theta = 0.8$ (right).

The EI has the following relation to the distribution of the maximum:

$$\mathbb{P}\{M_n \leq u_n\} = \mathbb{P}^{n\theta}\{X_1 \leq u_n\} + o(1) = F^{n\theta}(u_n) + o(1), \quad n \to \infty. \tag{3}$$

It holds $\theta = 1$ if the r.v.s $\{X_n\}$ are i.i.d.. The converse is incorrect. An EI that is close to zero implies a kind of a strong dependence. Stochastic processes with a strong local dependence and $\theta = 0$ exist. A Lindley process that models the waiting times in a G/GI/1 queueing system may provide such an example in case of a sub-exponentially distributed noise term, [2]. Relation (3) implies in the case $\theta = 0$ that the maximum will likely not exceed a sufficiently high threshold u_n, i.e. $\mathbb{P}\{M_n \leq u_n\} \to 1$ holds whenever u_n satisfies the first limit in (2).

In order to use results in [18], we assume that the p-h delays $\{X_{i,j} : i \geq 1\}$ at links with the number $1 \leq j \leq L_i$ of paths with numbers $i \geq 1$ are stationary distributed as in (1) and have their TI $k_j > 0$ and EI $\theta_j \in [0, 1]$, and that among

all sets of the links there exists a unique set with the minimal TI. Without loss of generality, this can be the set $\{X_{i,1} : i \geq 1\}$ of first links from the source node S with a TI equal to k_1. Such set of links is in a strong-sense stationary distributed with the heaviest distribution tail. Other sets have TIs larger than k_1 and, hence, according to (1) they are not so heavy-tailed distributed. Although some of such $\{k_j\}_{j\geq 2}$ may be equal, the corresponding distributions of the link sets may not be the same if the slowly varying functions $\ell_j(x)$, $j \geq 2$ in (1) are different. An arbitrary dependence between $\{X_{i,j}\}$ and L_i is allowed therein.

In [18] the EI and the TI of sums and maxima of random sequences of random lengths $\{L_n\}$ were considered. One can model the distribution tail of L_n as $\mathbb{P}\{L_n > x\} = \widetilde{\ell}_n(x)x^{-\alpha}$ with the TI $\alpha > 0$. Indeed, the lengths are integer-valued r.v.s. The relevance of such modeling is pointed out in several papers, see [7], [8], [30] among those. Assuming that both the slowly varying functions $\{\ell_j(x)\}$ in (1) and $\{\widetilde{\ell}_n(x)\}$ are bounded uniformly by polynomial functions for sufficiently large x over all sets of links and all path lengths, and that L_n has a lighter tail than the most heavy-tailed distributed p-h delay $X_{n,1}$, i.e. $\alpha > k_1$ holds, it is proved in [18] that the sequences of sums and maxima

$$X_n(z, L_n) = z_1 X_{n,1} + z_2 X_{n,2} + ... + z_{L_n} X_{n,L_n},$$
$$X_n^*(z, L_n) = \max(z_1 X_{n,1}, z_2 X_{n,2}, ..., z_{L_n} X_{n,L_n})$$

with positive constants $z_1, ..., z_{L_n}$ follow a distribution (1) with the same k_1 and θ_1. As the E2E delays constitute random sums of a random number of terms, the mentioned result relates to our problem. According to [15, Theorem 1], L_n is geometrically distributed irrespective of the distributions of the packet transmission rates and E2E delays and depending only on the levels of their quantiles. It is assumed that the per-hop transmission rates of the packets are i.i.d. and independent of the E2E transfer delay. Hence, the geometric model meets the result in [18], but L_n is assumed to be regularly varying distributed with a positive TI. The latter assumption is not restrictive since the class of distributions with regularly varying tails is rather wide.

In case that some paths include a node with light-tailed distributed p-h delay and(/or) the distribution of the p-h delays at some link from the source contains a mixture of light- and heavy- tailed distributions, the basic statistical result developed in [18] is still valid. This property follows from the proofs of Theorem 3 and 4 in [18].

3 Statistical Analysis of the End-to-End Delay

3.1 Asymptotic Distribution of the E2E and Maximal P-h Delays

Let us consider a path of random length $L_i(D)$ between the source and destination nodes (S, D). $L_i(D)$ is equal to the number of links between the source S and destination D. Let $n \geq 1$ be the number of possible paths constructed by the nodes of the P2P overlay network. Since the P2P network may be changed

dynamically in time, the number of nodes available for the packet transmission is changing and n is random. We can neglect its randomness considering the approach as a conditional one, since n is proportional to the number of nodes N in the network and the latter can be large. The theoretical result in [18] assumes that n is deterministic and tends to infinity.

Let us consider the double-indexed array of the p-h delays $\mathcal{X} = (X_{i,j} : i, j \geq 1)$. The "row index" i corresponds to the p-h delays belonging to the same path i between the source S and destination D, and the "column index" j corresponds to the p-h delays arising at the jth link enumerated from the source node. All p-h delays relate to the same source-destination pair (S, D). We consider the corresponding matrix

$$\mathcal{X} = \begin{pmatrix} X_{1,1} & X_{1,2} & 0 & ...0 & X_{1,L_1} \\ X_{2,1} & X_{2,2} & X_{2,3} & ...0 & X_{2,L_2} \\ ... & ... & ... & ... & ... \\ X_{n,1} & X_{n,2} & X_{n,3} & ...X_{n,L_n-1} & X_{n,L_n} \\ \begin{pmatrix} k_1, & k_2, & k_3, ..., & k_{L_n-1}, & k_{L_n} \\ \theta_1, & \theta_2, & \theta_3, ..., & \theta_{L_n-1}, & \theta_{L_n} \end{pmatrix} \end{pmatrix} \tag{4}$$

where the first and last columns corresponding to the one-hop links to the source and destination nodes are full and the internal columns are completed by zeros up to the maximal dimension $L_{max} = \max\{L_1, ..., L_n\}$, let's say $L_n = L_{max}$. We assume the most general case: the columns can be dependent, and each column may consist of dependent p-h delays, and the distribution of each column is stationary with the positive TI value k_j. Its local dependence structure is described by the EI value θ_j.

For any location of zeros in the matrix $\mathcal{X}$, the minimal TI (and the corresponding EI) of the internal columns taken together is determined by the distribution of the most heavy-tailed distributed element. This property follows from the proof of Theorem 3 in [18]. The sum $D_i = \sum_{j=1}^{L_i} z_j X_{i,j}$ and maximum $M_i = \max_{j=1,...,L_i}\{z_j X_{i,j}\}$ of weighted elements of the ith string set determine the weighted E2E delay between the source and destination nodes of the ith path and the longest weighted p-h delay at the ith path, respectively. The weights $\{z_i\}$ may reflect a priority which can be proportional to the capacities of links or impact on the scheduling of the peer selection process determining the path. In the simplest case, $\{z_i\}$ are all equal to one.

We suppose, without loss of generality, that the minimal TI value k_1 belongs to the first column and $k_1 < k$, with $k = \lim_{n\to\infty} \inf_{2\leq j\leq l_n} k_j$, $\quad l_n = \lceil n^\chi \rceil$, $0 < \chi < (k - k_1)/(k_1(k + 1))$ holds. The value k_1 corresponds to the heaviest distribution tail among the columns. According to Theorem 4 in [18] it follows

$$\mathbb{P}\{M_i > x\} = \mathbb{P}\{D_i > x\}(1 + o(1)) = \ell_1(x)x^{-k_1}(1 + o(1)) \tag{5}$$

as $x \to \infty$. This result means that the most heavy-tailed distributed column of the p-h delays determines the distributions of the E2E delay and the maximal p-h delay at the ith path. Instead of the E2E delays, one can consider the maximal p-h delays at each path (or vice versa) since they have the same heaviness of

tail, i.e. the same distribution up to the slowly varying functions. This allows us to model the distribution of the p-h delays since the E2E delays can be easily gathered as statistics in practice, rather than the p-h delays.

By Theorem 4 in [18] the EI of M_i and D_i is equal to the value θ_1 corresponding to k_1. Then the maxima of the sequences $\{D_i\}$ and $\{M_i\}$, $i = 1, ..., n$, have the same limiting distributions. More exactly, it holds

$$\lim_{n\to\infty} \mathbb{P}\{M_n^s \le u_n\} = \lim_{n\to\infty} \mathbb{P}\{M_n^m \le u_n\} = e^{-\tau\theta_1} \tag{6}$$

by (2) with

$$\lim_{n\to\infty} n\mathbb{P}\{M_n > u_n\} = \lim_{n\to\infty} n\mathbb{P}\{D_n > u_n\} = \tau, \tag{7}$$

where we denote

$$M_n^s = \max\{D_1, ..., D_n\}, \quad M_n^m = \max\{M_1, ..., M_n\},$$

and $\{u_n\}$ is an increasing sequence of thresholds. In [18] u_n is selected by (5) and (7) in such a way that $\tau = (z_1/y)^{k_1}$ with a constant $y > 0$ holds, namely, $u_n = yn^{1/k_1}\ell_1^\sharp(n)$, where $\ell_1^\sharp(n)$ is a slowly varying function.

Regarding the transmission rates of the packet flows we can argue in the same way. Following [15], each node is a bottleneck and it may upload an own superimposed flow coming from other nodes. Then a transported packet is associated with the sequence of transmission rates $\{R_{i,1}, R_{i,2}, ..., R_{i,L_i}\}$ corresponding to the links of the ith path. We approximate these transmission rates as ratios $R_{i,j} = Y_i/Z_{i,j}$, where Y_i is the packet length and $Z_{i,j}$ is the inter-arrival time between the considered packet and the previous (or next) one arriving at the jth node. Clearly, the rates $\{R_{i,j}\}$, $j = 1, 2, \ldots$ are all dependent for a fixed i.

Considering the matrix $\mathcal{X}$ in (4) one can substitute $X_{i,j}$ by $R_{i,j}$ assuming that the columns of the transmission rates have the TIs $\{k_i^*\}$ and EIs $\{\theta_i^*\}$ and that a unique minimal TI k_1^* exists as for the p-h delays:

$$\mathcal{R} = \begin{pmatrix} R_{1,1} & R_{1,2} & 0 & ...0 & R_{1,L_1} \\ R_{2,1} & R_{2,2} & R_{2,3} & ...0 & R_{2,L_2} \\ ... & ... & ... & ... & ... \\ R_{n,1} & R_{n,2} & R_{n,3} & ...R_{n,L_n-1} & R_{n,L_n} \end{pmatrix} \begin{pmatrix} k_1^*, & k_2^*, & k_3^*, ..., & k_{L_n-1}^*, & k_{L_n}^* \\ \theta_1^*, & \theta_2^*, & \theta_3^*, ..., & \theta_{L_n-1}^*, & \theta_{L_n}^* \end{pmatrix}. \tag{8}$$

Then we obtain (6) with corresponding replacements. As the result stated in [18] concerns weighted sums and maxima, one can think that some links may have a priority which can be proportional to their capacities or that the weights $\{z_i\}$ can impact on the scheduling of the peer selection process determining the path.

The probability of the successful transmission P_{st} of n packets over their n paths is determined by

$$P_{st} = \mathbb{P}\{M_n^{m*} \le u_n^*\} + \mathbb{P}\{M_n^m \le b_n\}, \tag{9}$$

where

$$M_n^{m*} = \max\{M_1^*, ..., M_n^*\}, \quad M_i^* = \max_{j=1,...,L_i} \{z_j R_{i,j}\}$$

are the maximal transmission rates of the packets over n paths and over the path i, respectively. The excess of the rate over the equivalent channel capacity u_n^* may cause the miss of packets [15]. In (9) $\mathbb{P}\{M_n^m \le b_n\}$ is the probability that the longest (weighted) p-h delay M_n over n paths is less than the playback delay b_n. The sequences $\{u_n^*\}$ and $\{b_n\}$ are determined to be increasing as $n \to \infty$ in the same way as $\{u_n\}$ in [18], i.e.

$$u_n^* = yn^{1/k_1^*}, \qquad b_n = yn^{1/k_1} \tag{10}$$

omitting the slowly varying functions for simplicity. Such sequences correspond to high quantiles of the rates and p-h delays. Then it holds

$$P_{st}(y) \approx e^{-\tau^*\theta_1^*} + e^{-\tau\theta_1} = e^{-(z_1/y)^{k_1^*}\theta_1^*} + e^{-(z_1/y)^{k_1}\theta_1}, \ y > 0 \tag{11}$$

for sufficiently large n, where y is selected in such a way to keep $P_{st}(y) < 1$. Hence, the approximate probability to loose at least one packet during the transmission over n paths is given by

$$P_m(y) = 1 - P_{st}(y). \tag{12}$$

Taking $P_m(y) = \eta$, where $\eta \in (0, 1)$ is small, one can find a corresponding y. In Fig. 3 an example is shown where $y = 0.755$ provides the solution to $P_m(y) = 0.05$.

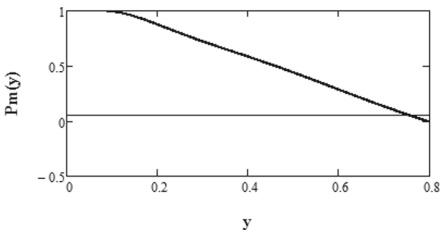

Fig. 3. P_m against y for $\alpha_1 = 1.2$, $\alpha_1^* = 2$, $\theta_1 = 0.3$, $\theta_1^* = 0.7$ and $z_1 = 1$ (thick solid line), $\eta = 0.05$ (thin solid line).

One may also consider a simple example of such calculation. Let us suppose that $e^{-\tau^*\theta_1^*} = e^{-\theta_1}$ holds. Then we get $\tau^* = \theta_1/\theta_1^*$, $1 - e^{-\theta_1} - e^{-\tau\theta_1} = \eta > 0$ and $\eta < 1 - e^{-\theta_1}$. Taking $\tau = (z_1/y)^{\alpha_1}$, we obtain

$$y = z_1 \exp\left(-\frac{1}{\alpha_1} \ln\left(-\frac{1}{\theta_1} \ln(1 - e^{-\theta_1} - \eta)\right)\right). \tag{13}$$

For example, if $z_1 = 1$, $\theta_1 = 0.3$, $\theta_1^* = 0.7$ holds, we may take $\eta < 0.259$, $\tau^* = 0.429$. Given $\alpha_1 = 1.2$ and $\eta = 0.01$, we get $y = 0.221$. However, such y does not depend on the TI and EI of the transmission rates.

To prepare our further statistical estimation techniques, let us finally summarize our general assumptions about the proposed statistical model of the E2E and p-h delays regarding the packet transfer in a P2P network. We assume that

(i) the P2P overlay network may be dynamic;
(ii) the p-h packet delays at the different links of the same path may be arbitrary dependent;
(iii) the p-h packet delays at the different links located on the same distance from the source node and belonging to different source-destination paths are stationary distributed, but not necessarily independent;
(iv) the length of a path and the p-h delays at its links may be dependent;
(v) the distribution of the path length has a lighter tail than the p-h delays with the heaviest tail;
(vi) there exists a unique set of links located on the same distance from the source that has the heaviest distribution tail compared to other sets of links among the overlay paths.

We recall that the E2E delays are regularly varying (heavy-tailed) distributed which follows from (5). The assumption (v) is fulfilled since the normalized path length is geometrically distributed irrespective of the distributions of the transmission rates and E2E delays and depending only on the levels of their quantiles [15].

3.2 Nonparametric Statistical Estimation

In the previous section we have considered asymptotic statistical results when the number of paths in a P2P network tends to infinity. Now we consider the case of finite samples.

The important step of the approach is to detect whether the unique column of the matrix $\mathcal{X}$ in (4) or $\mathcal{R}$ in (8) with the smallest TI exists or not. For this purpose the discrimination tests of close distribution tails built by only higher order statistics can be used, [23,24]. The application of such a test to each pair of columns of $\mathcal{X}$ or $\mathcal{R}$ to discriminate the heaviest tail consistently may constitute a calculation problem that is out of scope of this paper. Here, this problem can be solved from another perspective.

Many proposed network architectures place nodes with large upload capacities close to the source [5]. Thus, one may expect the smallest capacities and transmission rates at the last link before the destination node. This property may lead to the heaviest distribution tail of the p-h delays or the transmission rates and the smallest TI at the last link. Thus, one can estimate and compare the TIs and EIs of the p-h delays or the rates at the internal part and the last column of the matrix $\mathcal{X}$ or $\mathcal{R}$, respectively, and find the minimal k_1 or k_1^*, respectively.

Estimation of the TI. Let $X^n = \{X_1, ..., X_n\}$ be a sample of r.v.s with cumulative distribution function (cdf) $F(x)$. These r.v.s could be the transmission rates $R_{i,j}$ or the p-h delays $X_{i,j}$.

The Hill's estimator is well known and the simplest one to estimate the TI, but it requires an i.i.d. sample, [9], [14], [20], [29]. Regarding dependent data one can recommend estimators based on sums and maxima of non-intersecting data blocks, [19], [22]. The reduced bias estimator of the extreme value index that is the reciprocal of the TI is proposed in [4].

Several nonparametric estimators of the TI can be written by means of the statistic proposed in [21]

$$G_n(k, r, v) = \frac{1}{k} \sum_{i=0}^{k-1} g_{r,v} \left(\frac{X_{n-i,n}}{X_{n-k,n}} \right), \quad g_{r,v}(x) := x^r \ln^v(x),$$

where $r \in R$, $v > -1$. For instance, this estimator includes the Hill's estimator

$$\gamma_n^{(H)}(k) = G_n(k, 0, 1) = \frac{1}{k} \sum_{i=0}^{k-1} \ln \left(\frac{X_{n-i,n}}{X_{n-k,n}} \right), \tag{14}$$

or the moment-ratio estimator

$$\hat{\gamma}_n^{(mr)}(k) = G_n(k, 0, 2) \left(2G_n(k, 0, 1) \right)^{-1} \tag{15}$$

proposed in [6] to estimate the extreme value index $\gamma = 1/\alpha$ which is the reciprocal of the TI α. Here, $1 \le k \le n-1$ is the number of the largest order statistics

$$X_{n-k,n} \le X_{n-k+1,n} \le \cdots \le X_{n,n}$$

of the sample $\{X_1, \ldots, X_n\}$ used for the estimation, and r is a tuning parameter. The statistics $G_n(k, r, v)$ are special cases of the statistics introduced in [25].

The choice of k constitutes another problem. The simplest visual method is given by the Hill plot $\{(k, \gamma_n^{(H)}(k)) : k = 1, \ldots, n-1\}$. Then the estimate of k is selected from the interval $[k_-, k_+]$ of stability of the function $\gamma_n^{(H)}(k)$, [14]. Alternatives could be the exceedance plot or a bootstrap method as well as an exact calculation of k and r as in [20].

Estimation of the EI. Among the nonparametric estimators of the EI, the blocks, runs and intervals estimator are the most popular ones, [3]. As the reciprocal of the EI approximates the mean cluster size, the estimators differ by the definition of the cluster of exceedances. Particularly, the cluster of the blocks estimator is a data block with at least one exceedance over a threshold u. The blocks and runs estimators require a tuning parameter and the threshold u whereas the intervals estimator needs only u, [11].

The intervals estimator is calculated by a specific sample $\{T_1(u)\}_{i=1}^{L}$ of the length $L = L(u) < n$ generated by the initial sample $X^n = \{X_1, ..., X_n\}$. Namely,

$$T_1(u) = \min\{j \geq 1 : M_{1,j} \leq u, X_{j+1} > u | X_1 > u\}$$

denotes the number of consecutive non-exceedances between two consecutive clusters of exceedances, where $M_{1,j} = \max\{X_2, ..., X_j\}$, $M_{1,1} = -\infty$ holds. Here the cluster of exceedances determines a set of consecutive exceedances of the underlying stochastic sequence over the threshold u between two consecutive non-exceedances. Then the intervals estimator is defined as

$$\hat{\theta}_n(u) = \begin{cases} \min(1, \hat{\theta}_n^1(u)), & \text{if } \max\{(T_1(u))_i : 1 \leq i \leq L-1\} \leq 2, \\ \min(1, \hat{\theta}_n^2(u)), & \text{if } \max\{(T_1(u))_i : 1 \leq i \leq L-1\} > 2, \end{cases} \tag{16}$$

where

$$\hat{\theta}_n^1(u) = \frac{2(\sum_{i=1}^{L-1}(T_1(u))_i)^2}{(L-1)\sum_{i=1}^{L-1}(T_1(u))_i^2}, \tag{17}$$

$$\hat{\theta}_n^2(u) = \frac{2(\sum_{i=1}^{L-1}((T_1(u))_i - 1))^2}{(L-1)\sum_{i=1}^{L-1}((T_1(u))_i - 1)((T_1(u))_i - 2)} \tag{18}$$

holds. Among the last achievements, one can mention the K-gaps estimator that improves the intervals estimator, [28]. In [12] one can find the IMT method to calculate an optimal pair (u, K) for the K-gaps estimator.

Usually, u is chosen among those quantiles that are higher than 95% of an underlying sequence. u can be selected visually as corresponding to the stability interval of the plot $\{(u, \widehat{\theta}(u))\}$ in the same way as the Hill plot. One can apply a bootstrap method [17] or the discrepancy method [16] for its automatic selection.

Then we can determine the basic nonparametric estimation algorithm by these statistical means.

Estimation Algorithm. Let us consider the last columns of the matrices $\mathcal{X}$ in (4) and $\mathcal{R}$ in (8), namely, $\{X_{i,L_i}\}$ and $\{R_{i,L_i}\}$, $i = 1, 2, ..., n$, as initial data. Here n is the number of possible paths between the source and destination nodes (S, D) of the P2P overlay network.

1. Estimate the TIs α_{L_n} and $\alpha^*_{L_n}$ by $\{X_{i,L_i}\}$ and $\{R_{i,L_i}\}$ using one of the nonparametric estimators, e.g. (14) or (15).
2. Estimate the EIs θ_{L_n} and $\theta^*_{L_n}$ by $\{X_{i,L_i}\}$ and $\{R_{i,L_i}\}$ using one of the nonparametric estimators, e.g. (16)–(18).
3. Calculate y as $y = \arg\{t : P_m(t) = \eta\}$ or by (13) for a predefined $0 < \eta < 1$.
4. Calculate the probabilities $P_{st}(y)$ and $P_m(y)$ by (11) and (12).
5. Calculate the equivalent capacity u_n^* and the playback delay b_n by (10).

3.3 An Illustrative Example

In this section our aim is to demonstrate the sketched methodology using simulated examples of sequences that are arising from regularly varying distributed r.v.s. We simulate samples of the transmission rates $\{R_{i,L_i}, i \in \{1,2,\ldots,n\}\}$ as Moving Maxima (MM) process and of the p-h delays $\{X_{i,L_i}, i \in \{1,2,\ldots,n\}\}$ as MA(2) process. Considering a P2P network in practice, indeed, the real processes could be different. However, our methodology is a pure nonparametric approach and can be applied to any process model.

The mth order MM process is determined by

$$X_t = \max_{i=0,\ldots,m} \{\beta_i Z_{t-i}\}, \qquad t \in \mathbb{Z},$$

where $\{\beta_i\}$ are constants with $\beta_i \geq 0$, $\sum_{i=0}^{m} \beta_i = 1$, and Z_t are i.i.d. standard Fréchet distributed r.v.s with the cdf $F(x) = \exp(-1/x)$ for $x > 0$. The EI of the process is equal to $\theta = \max_i\{\beta_i\}$ [1]. The distribution of $\{X_t\}_{t\geq 1}$ is standard Fréchet. Its TI is equal to one. In our study the values $m = 3$ and $\theta = 0.5$ corresponding to $\beta \in \{0.5, 0.3, 0.15, 0.05\}$ are selected.

The MA(2) process is determined by

$$X_i = pZ_{i-2} + qZ_{i-1} + Z_i, \qquad i \geq 1, \tag{19}$$

with $p > 0$, $q < 1$, and i.i.d. Pareto random variables $Z_{-1}, Z_0, Z_1, \ldots$ with $\mathbb{P}\{Z_0 > x\} = 1$ if $x < 1$, and $\mathbb{P}\{Z_0 > x\} = x^{-\alpha}$ if $x \geq 1$ hold for some $\alpha > 0$ [27]. The EI of the process is given by $\theta = (1 + p^\alpha + q^\alpha)^{-1}$. The case $\alpha = 2$, $(p, q) = (1/\sqrt{2}, 1/\sqrt{2})$ with a corresponding value $\theta = 0.5$ is considered. Since the distribution of the sum of weighted i.i.d. Pareto r.v.s behaves like a Pareto distribution in the tail, namely,

$$\mathbb{P}\{\sum_{i=1}^{n} Z_i > x\} \sim n(1 + x/\beta)^{-\alpha} \cdot L(x), \quad x \to \infty, \tag{20}$$

where $L(x)$ is a slowly varying function at infinity, $\beta > 0$ is a scale parameter, and $\alpha > 0$ is the TI, (see [10, Ch. 8, pp. 268–272]), then X_i is also Pareto distributed with the TI α.

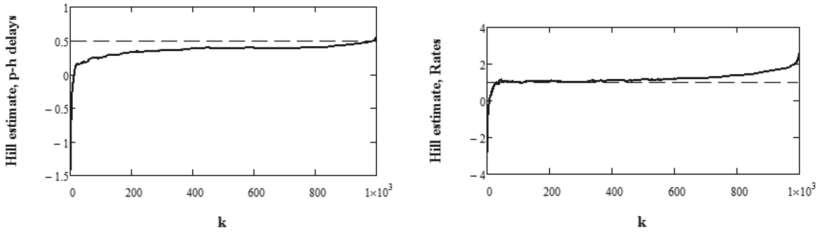

Fig. 4. The Hill's estimate of the p-h delays modeled as MA(2) process with the TI $\alpha = 2$ (and the EVI $\gamma = 0.5$) (lhs) and the transmission rates modeled as MM process with the TI $\alpha = 1$ (and $\gamma = 1$) (rhs); the sample size is given by $n = 1000$.

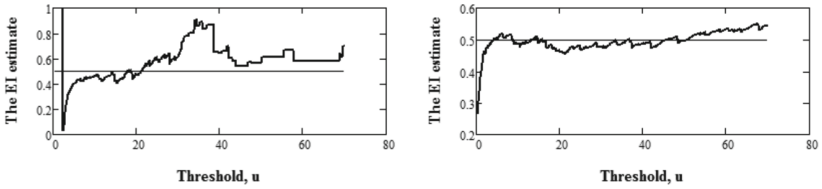

Fig. 5. The intervals estimate of the p-h delays modeled as MA(2) process with the EI $\theta = 0.5$ (lhs) and the transmission rates modeled as MM process with the EI $\theta = 0.5$ (rhs); the sample size is given by $n = 20000$.

Our objective is to estimate the parameters and to calculate all relevant metrics according to the proposed estimation algorithm using these simulated samples.

The Hill's estimator is very sensitive to the presence of a slowly varying function in the distribution tail. Thus, the estimate on the left-hand side of Fig. 4 corresponding to (19) and (20) is rather biased. In practice it is therefore reasonable to use several estimators of the TI.

Now we consider the EI estimation by the intervals estimator (16)–(18) that is applied to the same processes MA(2) and MM, see Fig. 5. Here we have to generate larger samples with $n = 20000$. The intervals estimator requires a large sample size of $\{X^n\}$ to get a better estimation since it is based on the sample of the inter-exceedance times $\{T_1(u)\}_i$, $i = 1, 2, ..., L(u)$, generated from the underlying sample X^n. The size of $\{T_1(u)\}_i$ can be much smaller than n depending on the threshold u, the higher u the smaller $L(u)$.

We can obtain $y \in \{0.768, 0.732\}$ for a given $\eta \in \{0.05, 0.1\}$, respectively, and for given $\alpha_1^* = 1$, $\alpha_1 = 2$, $\theta_1^* = \theta_1 = 0.5$ in the same way as in Fig. 3. By formulae (10) we then obtain for $n = 1000$ $u_n^* \in \{768, 732\}$ and $b_n \in \{24.286, 23.148\}$, respectively. Regarding such y the probabilities $P_{st}(y)$ and $P_m(y)$ calculated by (11) and (12) are equal to $\{0.95, 0.898\}$ and $\{0.05, 0.102\}$, respectively. We note that the maximal values of the generated random sequences $\max_{1 \le i \le n} X_i$ are equal to 26.946 w.r.t. the p-h delays and 743.439 w.r.t. the transmission rates. It implies that u_n^* and b_n exceed these maxima, and $P_{st} = 0.95$ is not realistic for these models. A calculation of y by (13) provides $u_n^* = 522$ and $b_n = 16.507$. Such low thresholds immediately reflect on P_{st} and P_m providing $P_{st} = 0.543$ and $P_m = 0.457$, respectively.

4 Conclusions and Open Problems

We have considered the performance analysis of the data transfer along transport paths of random lengths in a P2P overlay network subject to QoS constraints. First, the distribution of the end-to-end (E2E) transfer delay of the packet flows between the source and destination nodes is modeled. The E2E transfer delay is determined by the sum of a random number of p-h delays along the links of an overlay path. Based on recent statistical results in [18] and assuming that the per

hop (p-h) delays and the lengths of the paths are regularly varying distributed, it is shown that the sums and maxima of the p-h delays corresponding to different paths of random lengths may have the same tail and extremal indexes TI and EI, respectively. These indexes determine the heaviness of the tail of the delay distribution and the dependence indicator that measures the cluster tendency (i.e., how extreme values arise by groups of observations). Using the EI, then the limit distributions of the maxima of the E2E and p-h delays over all source-destination paths are identified. Considering real-time applications with stringent E2E delay constraints, the latter distributions are used to identify important QoS metrics of a P2P-model like the packet missing probability, the corresponding playback delay, and the required equivalent capacity to transfer the packet flows of the data.

The proposed approach requires the verification and comparison of the TIs of the p-h delays to find the set of links whose delays have the heaviest tail. Regarding modern network architectures one can expect that the last link before the destination node has the heaviest distribution tail. Then known statistical tests allow us to compare pairs of samples in the columns of the matrix $\mathcal{X}$ (or $\mathcal{R}$) regarding the similarity of their distributions. We note that the lengths of the overlay paths of packet flows in a P2P network can be observed if the packet header is providing a counter of the visited nodes along the path. Then the TI of the lengths can be estimated by these means.

The described asymptotic results are valid for sufficiently high thresholds that are in our context the playback delay and the equivalent capacity of the transport channel. Our statistical results provide the basis for an improved control scheme regarding the optimal selection of transport paths in a P2P overlay network subject to QoS constraints on the E2E delay and packet loss metrics.

Regarding the application of a P2P overlay concept in 5G networks, we may look at the deployment of a blockchain functionality on top of an underlying network of mining peers that are validating transactions of IoT data processing or the use of P2P video streaming as important examples. In the case of such real-time applications, we are looking for short playback delays, but they may lead to a large packet missing probability. In this respect the derived asymptotic performance analysis models of the E2E transfer delay provide a tendency with an increasing probability of successful packet transmission as both the playback delay and the equivalent capacity increase. But these performance analysis models require an adjustment for short playback delays and not high, realistic capacities.

Our future studies will focus on these analysis and design issues of modern teletraffic theory.

Acknowledgments. The first author was partly supported by Russian Foundation for Basic Research (grant 19-01-00090).

References

1. Ancona-Navarrete, M.A., Tawn, J.A.: A comparison of methods for estimating the extremal index. Extremes **3**(1), 5–38 (2000). https://doi.org/10.1023/A:1009993419559
2. Asmussen, S.: Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities. Ann. Appl. Probab. **8**, 354–374 (1998)
3. Beirlant, J., Goegebeur, Y., Teugels, J., Segers, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
4. Caeiro, F., Gomes, M.I., Beirlant, J., de Wet, T.: Mean-of-order p reduced-bias extreme value index estimation under a third-order framework. Extremes **19**, 561–589 (2016). https://doi.org/10.1007/s10687-016-0261-5
5. Dán, G., Fodor, V.: Delay asymptotics and scalability for peer-to-peer live streaming. IEEE Trans. Parallel Distrib. **20**(10), 1499–1511 (2009)
6. Danielsson, J., Jansen, D.W., de Vries, C.G.: The method of moments ratio estimator for the tail shape parameter. Commun. Stat. Theory. **25**, 711–720 (1986)
7. Jelenkovic, P.R., Olvera-Cravioto, M.: Information ranking and power laws on trees. Adv. Appl. Prob. **42**(4), 1057–1093 (2010)
8. Jessen, A.H., Mikosch, T.: Regularly varying functions. Publ. Inst. Math. (Beograd) (N.S.) **80**, 171–192 (2006)
9. Hill, B.M.: A simple general approach to inference about the tail of a distribution. Ann. Stat. **3**, 1163–1174 (1975)
10. Feller, W.: An Introduction to Probability and Its Application, 2nd edn. Wiley, New York (1971)
11. Ferro, C.A.T., Segers, J.: Inference for clusters of extreme values. J. R. Stat. Soc. B. **65**, 545–556 (2003)
12. Fukutome, S., Liniger, M.A., Süveges, M.: Automatic threshold and run parameter selection: a climatology for extreme hourly precipitation in Switzerland. Theoret. Appl. Climatol. **120**, 403–416 (2015)
13. Leadbetter, M.R., Lingren, G., Rootzen, H.: Extremes and Related Properties of Random Sequence and Processes. Chap. 3. Springer, New York. https://doi.org/10.1007/978-1-4612-5449-2 (1983)
14. Markovich, N.M.: Nonparametric Estimation of Univariate Heavy-Tailed Data. Wiley, Chichester (2007)
15. Markovich, N.M.: Quality assessment of the packet transport of peer-to-peer video traffic in high-speed networks. Perform. Eval. **70**, 28–44 (2013)
16. Markovich, N. M.: Nonparametric estimation of extremal index using discrepancy method. In: Proceedings of the X International Conference "System Identification and Control Problems" SICPRO-2015, Moscow, V.A. Trapeznikov Institute of Control Sciences, 26–29 January, pp. 160–168 (2015)
17. Markovich, N.M., Ryzhov, M.S., Krieger, U.R.: Statistical clustering of a random network by extremal properties. In: Vishnevskiy, V.M., Kozyrev, D.V. (eds.) DCCN 2018. CCIS, vol. 919, pp. 71–82. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99447-5_7
18. Markovich, N.M., Rodionov, I.V.: Maxima and sums of non-stationary random length sequences. Extremes **23**(3), 451–464 (2020). https://doi.org/10.1007/s10687-020-00372-5

19. Markovich, N., Vaičiulis, M.: Modification of moment-based tail index estimator: sums versus maxima. In: Bertail, P., Blanke, D., Cornillon, P.-A., Matzner-Løber, E. (eds.) ISNPS 2016. SPMS, vol. 250, pp. 85–101. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-96941-1_6
20. Vaičiulis, M., Markovich, N.M.: A class of semiparametric tail index estimators and its applications. Autom. Remote Control **80**(10), 1803–1816 (2019). https://doi.org/10.1134/S0005117919100035
21. Paulauskas, V., Vaičiulis, M.: Several new tail index estimators. Ann. Inst. Stat. Math. **69**, 461–487 (2017)
22. McElroy, T., Politis, D.N.: Moment-based tail index estimation. J. Statist. Plan. Infer. **137**, 1389–1406 (2007)
23. Rodionov, I.V.: On discrimination between classes of distribution tails. Probl. Inform. Transm. **54**(2), 124–138 (2018)
24. Rodionov, I.V.: Discrimination of close hypotheses about the distribution tails using higher order statistics. Theory Probab. Appl. **63**(3), 364–380 (2019)
25. Segers, J.: Residual estimators. J. Stat. Plan. Inf. **98**, 15–27 (2001)
26. Shih, M.F., Hero, A.O.: Unicast-based inference of network link delay distributions using mixed finite mixture models. IEEE Trans. Signal Process. **51**(8), 2219–2228 (2003)
27. Sun, J., Samorodnitsky, G.: Multiple thresholds in extremal parameter estimation. Extremes **22**, 317–341 (2019). https://doi.org/10.1007/s10687-018-0337-5
28. Süveges, M., Davison, A.C.: Model misspecification in peaks over threshold analysis. Ann. Appl. Statist. **4**(1), 203–221 (2010)
29. Vaičiulis, M.: Local-maximum-based tail index estimator. Lith. Math. J. **54**(4), 503–526 (2014)
30. Volkovich, Y.V., Litvak, N.: Asymptotic analysis for personalized web search. Adv. Appl. Prob. **42**(2), 577–604 (2010)