MODELING END-USER BEHAVIOR IN DATA NETWORKS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Luis López-Oliveros

August 2011

MODELING END-USER BEHAVIOR IN DATA NETWORKS

Luis López-Oliveros, Ph.D.

Cornell University 2011

A session is a cluster of packets that represents end-user activity in data networks, e.g. surfing the web, transferring files, streaming media, Internet-calling. We study three problems of sessions.

The first problem examines why for real sessions the distribution of the number of bytes or number of packets per unit time traveling through a network node is approximately Gaussian, despite previous theoretical results say it may also be stable Lévy. The second and third problems study four key session features: size or number of bytes transmitted, duration, average transfer rate, and initiation time. We focus on marginal distributions of size, duration and rate, dependence structure between the marginals, and distribution of the difference between consecutive initiation times. We group sessions according to peak transfer rate in the second problem, and network application in the third problem. The ultimate goal is to teach computers how to mimic network sessions and understand network end-user behavior.

BIOGRAPHICAL SKETCH

Luis López-Oliveros was born on June 6, 1981 in Santiago Tuxtla, Veracruz, México. After 14 years of moving between cities, his family settled down in Xalapa, where he graduated from La Oficial B in 1999. Luis received a bachelor degree in pure mathematics from the University of Guanajuato and the Center for Mathematical Research (CIMAT) in 2004. In CIMAT, he also received a masters degree in probability and statistics.

Luis joined the Department of Statistical Science of Cornell University in 2005. On August 16, 2008, he married his long time girlfriend Jazmin Becerra-Diez. After defending his dissertation on June 21, 2011, he joined Murex, North America, in New York City.

ACKNOWLEDGEMENTS

Professors Sidney Resnick, Gennady Samorodnitsky and Robert Jarrow comprised my special committee.

I could not have had a better research guide than Prof. Resnick. I learned a lot of extreme value and probability theory from him in our weekly meetings and in his courses. His ability with data analysis enhanced this work. His comments made this thesis much clearer. I much appreciated his encouragement, support, and friendship.

Prof. Samorodnitsky's course on long range dependence helped me understand many network traffic features, and his courses suggestions provided me with a well-balanced knowledge of applied probability.

Prof. Jarrow motivated my interest in financial markets with his course on fixed income and his course suggestions in financial engineering. His comments in my A and B exams fueled my interest in possible future applications of my research in other fields different than data networks.

Prof. Shane Henderson acted as a proxy special committee member; I appreciated his comments about my talk regarding Chapter 2. Stefan Weber also acted as a proxy special committee member at my A-exam.

Early discussions with Janet Heffernan significantly shaped the directions of the research in Chapter 3 and the final effort reflects the benefit of her initial creative inputs; in particular the idea of using link functions in Section 3.4.4 was hers. Edgar Bernal's suggestions helped build algorithms for harvesting sessions from packet headers in Chapters 3 and 4. Cornell Information Technologies-Network Communication Services' Dan Eckstrom and Ed Kiefer were very helpful with the collection of the Cornell flow data in Chapter 2. The Réseaux IP Européens Network Coordination Centre provided the data set in

Chapter 3. The Cooperative Association for Internet Data Analysis (CAIDA) provided Internet traces in Chapter 4; in particular, CAIDA's kc claffy, Dan Andersen and Paul Hick clarified traces content and format. Cornell Operations Research and Information Engineering's Eric Johnson and Department of Statistical Science's Todd Cullen were helpful with arrangements for data storage and software tools. Prof. Giles Hooker's comments partially shaped the presentation for my B exam.

The Department of Statistical Science provided financial support for four years. I am specially grateful with Prof. Robert Strawderman for finding me financial sources during my last year at Cornell. I also was supported for two years by the Mexican Research Council of Science and Technology (CONACyT) Contract 161069. S. While working with Prof. Resnick, he was partially supported by the Army Research Office Contract W911NF-07-1-0078 at Cornell University. Support for CAIDA's Internet Traces was provided by the National Science Foundation, the US Department of Homeland Security, and CAIDA Members.

Víctor Pérez-Abreu, Eloisa Díaz-Francés, Miguel Nakamura and David Sprott encouraged me to get a doctoral degree. Prior to coming to Cornell, they all were the pillars of my mathematical education. Víctor and Miguel have been good friends, and Eloisa and David motivated my interest in statistical applications. They all were excellent advisors.

My wife Jazmin burned the midnight oil with me for so many times. She also made sure that I had enough food and drink, and plenty of love. My parents Isabel and Luis, and my sister Mina have been there for me every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\bar{F}$	The right tail of the distribution function F, i.e. $\bar{F} = 1 - F$.
$F^{\leftarrow}$	The left continuous inverse of the distribution function F, i.e. $F^{\leftarrow}(y) = \inf\{x : F(x) \geq y\}$.
$f_1 \sim f_2$	$\lim_{x\to\infty} f_1(x)/f_2(x) = 1$.
$\xrightarrow{fidi}$	Convergence of finite dimensional distributions.
$\xrightarrow{v}$	Vague convergence of measures. See e.g. Kallenberg (1983); Resnick (1987).
$M_+(0,\infty]$	The space of nonnegative Radon measures on $(0,\infty]$.
RV_α	The class of regularly varying functions with index α. See e.g. Bingham et al. (1987).
$\mathcal{D}(G_\gamma)$	The maximal domain of attraction of the extreme value distribution G_γ. See de Haan and Ferreira (2006); Resnick (1986).
$\xi = \mathbb{PRM}(E\xi)$	A Poisson random measure ξ with mean measure $E\xi$.
$\mathring{\xi}$	A compensated Poisson random measure with mean measure $E\xi$, i.e. $\mathring{\xi} = \xi - E\xi$.
$N^{\infty}_{\alpha,h_\delta}$	$N^{\infty}_{\alpha,h_\delta} = \mathbb{PRM}(ds \cdot \alpha u^{-(\alpha+1)}du \cdot h_\delta(dr))$ on $\mathbb{R}\times(0,\infty)^2$, where h_δ is a measure on $(0,\infty)$ such that $h_\delta[\cdot,\infty) \in RV_{-\delta}$. If $h_\delta(dr) = \delta r^{-(\delta+1)}dr$, we simply write $N^{\infty}_{\alpha,\delta}$.
$M_{\alpha,m}(dv)$	A $\alpha-$stable random measure with control measure $m(dv)$ and stable index $1<\alpha<2$. For $\xi = \mathbb{PRM}(m(dv)w^{-(\alpha+1)}dw)$, we can write $M_{\alpha,m}(A) \stackrel{d}{=} \left(\left(-\cos\frac{\pi\alpha}{2}\right)\frac{2\Gamma(2-\alpha)}{\alpha(\alpha-1)}\right)^{-1/\alpha}\int_A\int_{w=0}^{\infty} w\mathring{\xi}(dv,dw)$. See e.g. Samorodnitsky and Taqqu (1994, Chapter 3).
$\Lambda_\alpha(\cdot)$	A $\alpha-$stable Lévy motion totally skewed to the right with stable index $1<\alpha<2$. In general, we can write $\Lambda_\alpha(t) \stackrel{d}{=} \left(\left(-\cos\frac{\pi\alpha}{2}\right)\frac{2\Gamma(2-\alpha)}{\alpha(\alpha-1)}\right)^{1/\alpha}\int_0^{\infty} 1_{\{0<v<t\}}M_{\alpha,m}(dv)$. See e.g. Samorodnitsky and Taqqu (1994, Chapter 3).
$B_H(\cdot)$	The standard fractional Brownian motion with Hurst exponent H.

CHAPTER 1
INTRODUCTION

Statistics on data networks show empirical features that are surprising by the standards of classical queuing theory. Three distinctive properties, which are called *invariants* in the network literature, are:

- Heavy tails for quantities such as file sizes (Leland et al., 1994; Willinger et al., 1998; Arlitt and Williamson, 1996; Willinger and Paxson, 1998), transmission durations and transmission delays (Maulik et al., 2002; Resnick, 2003).
- Network traffic is bursty (Sarvotham et al., 2005), with rare but influential periods of high transmission rate punctuating typical periods of modest activity. Burstiness is a somewhat vague concept but it is very important in order to understand network congestion.
- Gaussian cumulative traffic is seen in a heavily loaded network link subject to aggregation over many users (Leland et al., 1994; Kurtz, 1996; Willinger et al., 1997).

Here, we analyze various aspects of network end-user activity closely related to these invariants through end-user *sessions*. For now, think of a session as someone getting their email, surfing a website, downloading a music file, streaming a movie or a radio station, or skyping with friends and family. Our goal is to build mathematical models whose properties and predictions match empirical observations and to construct simulation methodology of sessions in order to understand end-user network activity, prevent congestion and identify bottlenecks.

In this chapter, we review basic concepts of network sessions and extreme value theory. We then introduce the problems covered in this dissertation.

1.1 Network sessions

Data networks like the Internet are called *packet-switched*. This means that transmissions over the Internet do not occur in a single piece, but rather in several small packets of data of bounded maximum size that depends on the specific network protocol. Thus, packet-level network traffic traces consist of records of packet headers, containing information of each individual packet such as arrival times to servers, number of bytes transmitted, source and destination network addresses, port numbers, transport protocols, etc. As the packets travel across the network, routers and switches use the packet header information to move each packet to its correct destination. The two main goals of packet-switching are to optimize the utilization of available line bandwidth and to increase the robustness of communication (see e.g. Keshav, 1997).

The nature of the network data sets poses a challenging question for modeling end-user activity: How do we reconstruct such activity from network packet headers?

We answer this question by clustering packets with the same source and destination network addresses according to some chosen but not unique rule. Various criteria for grouping packets yield different entities, e.g. connections, flows (or unidirectional connections), end-to-end streams, etc. (See e.g. Sarvotham et al., 2005, Section 4). These high-order constructs of packet clusters are sometimes termed *sessions*. Network administrators use different definitions of ses-

sions depending on their own goals. In the following chapters, we will precisely define the type of session we are dealing with, which depends partially on the problem of study and the available data sets.

Summary measurements are computed for

- S, the size, that is the number of bytes transmitted in the session.
- D, the duration of the session.
- R, the average transfer rate, namely S/D.
- Γ, the starting time of the session.

As pointed out in the introduction, data sets of S, D and R have historically shown heavy tails. We now make precise what we mean by that.

1.2 Heavy tails and maximal domains of attraction

A positive random variable Y has *heavy tails* if its distribution function F satisfies

$$1 - F(y) = \bar{F}(y) = y^{-1/\gamma} L(y), \tag{1.1}$$

where L is a slowly varying function and $\gamma > 0$. We also say that F is heavy tailed and we call γ the shape parameter. When F satisfies (1.1), it is also said to have regularly varying tails with tail index $\alpha = 1/\gamma$. The parameterization based on the shape parameter γ is useful for practical applications, while the parameterization based on the tail index α is better for theoretical developments. In this dissertation, we mostly use the former parameterization, but we will make a note wherever we use the latter one.

Equation (1.1) is equivalent to the existence of a sequence $b_n \to \infty$ such that

$$\mu_n(\cdot) := n\mathbb{P}\left[\frac{Y}{b_n} \in \cdot\right] \xrightarrow{v} c\nu_\gamma(\cdot), \tag{1.2}$$

vaguely in $M_+(0,\infty]$, the space of Radon measures on $(0,\infty]$. Here $\nu_\gamma(x,\infty] = x^{-1/\gamma}$ and $c > 0$. Equation (1.2) is important for generalizing the concept of heavy tailed distributions to higher dimensions.

An important concept is maximal domains of attraction. Suppose $\{Y_i; i \geq 1\}$ is iid with common distribution F. The distribution F is in the *maximal domain of attraction* of the extreme value distribution G_γ, denoted $F \in \mathcal{D}(G_\gamma)$, if there exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that for $y \in \mathbb{E}^{(\gamma)} = \{y \in \mathbb{R} : 1 + \gamma y > 0\}$:

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{\bigvee_{i=1}^n Y_i - b_n}{a_n} \leq y\right] = G_\gamma(y) := \exp\left\{-\left(1+\gamma y\right)^{-1/\gamma}\right\}. \tag{1.3}$$

This is equivalent to the existence of functions $a(t) > 0$ and $b(t) \in \mathbb{R}$ such that for $y \in \mathbb{E}^{(\gamma)}$:

$$\lim_{t\to\infty} t\mathbb{P}\left[Y_1 > a(t)y + b(t)\right] = -\ln G_\gamma(y). \tag{1.4}$$

The class of distributions $\mathcal{D}(G_\gamma)$ is known as the Fréchet domain when $\gamma > 0$, Gumbel domain when $\gamma = 0$ and Weibull domain when $\gamma < 0$. The limit in (1.3) is also known as the extreme value condition.

For $\gamma > 0$,

$$\bar{F}(y) = y^{-1/\gamma}L(y) \Leftrightarrow F \in \mathcal{D}(G_\gamma), \tag{1.5}$$

for some slowly varying L. In other words, a necessary and sufficient condition for a distribution to be heavy tailed is that it is in the Fréchet class (de Haan and Ferreira, 2006; Resnick, 1987).

One common method (Davison and Smith, 1990; Beirlant et al., 2004; Coles, 2001; Reiss and Thomas, 2007; Mc Neil et al., 2005; de Haan and Ferreira, 2006)

to check the extreme value condition, given by (1.3), relies on threshold excesses, using all data that are "extreme" in the sense that they exceed a particular designated high level. More precisely, consider a random variable Y with distribution function F. Given realizations of Y, say $y_1, \ldots, y_n$ and a threshold u, we call y_j an exceedance over u if $y_j > u$, and in such case, $y_j - u$ is called the *excess*. Denote the *excess distribution* over the threshold u as F_u, i.e.

$$F_u(y) = \mathbb{P}[Y - u \leq y | Y > u],$$

for all $0 \leq y \leq y_F - u$, where $y_F \leq \infty$ is the right endpoint of F. The connection with domains of attraction is that

$$F \in \mathcal{D}(G_\gamma) \Leftrightarrow \lim_{u \to y_F} \sup_{0 \leq y \leq y_F - u} \left| F_u(y) - GPD_{\gamma,\beta(u)}(y) \right| = 0, \text{ for some } \beta(u) > 0. \tag{1.6}$$

Here $GPD_{\gamma,\beta}$, with $\gamma \in \mathbb{R}, \beta > 0$ is the generalized Pareto distribution , defined as

$$GPD_{\gamma,\beta}(y) := 1 - (1 + \gamma y/\beta)^{-1/\gamma},$$

for $y \geq 0$ when $\gamma \geq 0$ and $0 \leq y \leq -\beta/\gamma$ when $\gamma < 0$. See Pickands (1975); Balkema and de Haan (1974); de Haan and Ferreira (2006).

For a distribution F, the method to check the extreme value condition using excesses over high thresholds (also referred to as peaks over thresholds or *POT*) assumes equality in (1.6) holds for a high threshold u, without need to take a limit, meaning that the excess distribution over such u equals a generalized Pareto distribution. See Embrechts et al. (1997); Coles (2001); Reiss and Thomas (2007); de Haan and Ferreira (2006). Suppose $Y_1, \ldots, Y_n$ are iid with common distribution F and let $Y_{1:n} \leq Y_{2:n} \leq \cdots \leq Y_{n:n}$ be the order statistics. Fix a high threshold $\hat{u} = Y_{n-k:n}$ as the $(k+1)$th largest statistic, and fit a $GPD_{\gamma,\beta}$ model to $Y_{n-k+1:n} - \hat{u}, \ldots, Y_{n:n} - \hat{u}$. Then the evidence supports $F \in \mathcal{D}(G_\gamma)$ if

and only if for some high threshold $\hat{u}$ that fit is adequate. For informally assessing the goodness of fit, we use QQ plots to compare sample quantiles, namely $\hat{Y}_{n-k+1:n} - \hat{u}, \ldots, \hat{Y}_{n:n} - \hat{u}$, against the theoretical quantiles given by the GPD fit. It is not difficult to show that $Z \sim GPD_{\gamma,\beta}$ is equivalent to the statement that $\ln(1 + \gamma Z/\beta))/\gamma \sim \exp(1)$, and so we draw QQ plots in this latter scale after estimating γ, β by means of, say, maximum likelihood. QQ plots of observations under this (*logarithmic*) transformation will be referred to as *the GPD* or *exponential QQ plots of the log-data* (or $\ln Z$ when we mention the specific data set).

Another popular estimator of γ is the Hill estimator (Hill, 1975; Csörgő et al., 1985; Davis and Resnick, 1984; de Haan and Resnick, 1998; Hall, 1982). The *Hill estimator* based on the k largest order statistics is

$$\hat{\gamma}_{k,n} = \frac{1}{k} \sum_{i=n-k+1}^{n} \ln \frac{Y_{i:n}}{Y_{n-k:n}}, \quad k = 1, \ldots, n-1. \tag{1.7}$$

For $F \in \mathcal{D}(G_\gamma), \gamma > 0$, the Hill estimator $\hat{\gamma}_{k,n}$ is a consistent estimator of γ. Furthermore, under an additional second order condition:

$$\sqrt{k}(\hat{\gamma}_{k,n} - \gamma) \xrightarrow{d} N(0, \gamma^2), \tag{1.8}$$

so usually both consistency and asymptotic normality hold as $k \to \infty, k/n \to 0$, and $n \to \infty$. See, for example, Hill (1975); Csörgő et al. (1985); Davis and Resnick (1984); de Haan and Resnick (1998),

The Hill estimator depends on the number k of upper-order statistics and so in practice, we make a *Hill plot* $\{(k, \hat{\gamma}_{k,n}); k \geq 1\}$ and pick a value of $\hat{\gamma}_{k,n}$ for which the graph looks stable. See e.g. Figure 3.4, de Haan and Ferreira (2006); Resnick (2007); Geluk et al. (1997); de Haan and Resnick (1998); Peng (1998); de Haan and Peng (1998); Mason and Turova (1994).

1.3 The infinite-source Poisson model

Our basic model for network transmission considers an infinite number of network nodes. Depending on whether we wear a mathematician or an empirical modeler hat, we assume or check that a homogeneous Poisson process on $\mathbb{R}$ with parameter λ activates data sessions at times $\{\Gamma_k, -\infty < k < \infty\}$. Each initiation time has an associated mark $\{(S_k, D_k, R_k) = (size, duration, rate), -\infty < k < \infty\}$.

We assume that the marks $\{(S_k, D_k, R_k)\}$ are independent and identically distributed, and independent of $\{\Gamma_k\}$. We will also assume or check that the marginal distributions of the triple are heavy-tailed.

1.4 Network problems of interest

We are interested in three problems of network end-user activity:

In Chapter 2 we study approximations of the distribution of cumulative network traffic, that is, the number of bytes or number of packets per unit time traveling through a network node. Various empirical and theoretical studies indicate that cumulative network traffic is a Gaussian process. However, depending on whether the intensity at which sessions are initiated is large or small relative to the session duration tail, Mikosch et al. (2002) and Kaj and Taqqu (2008) have shown that traffic at large time scales can be approximated by either fractional Brownian motion (fBm) or stable Lévy motion. We study distributional properties of cumulative traffic that consists of a finite number of independent streams and give an explanation of why Gaussian examples abound in practice

but not stable Lévy motion. We offer an explanation about how many streams are needed for the Gaussian approximation to hold. Our results are expressed as limit theorems for a sequence of cumulative traffic processes whose session initiation intensities satisfy growth rates similar to those used in Mikosch et al. (2002).

Whereas the problem in Chapter 2 is theoretical, Chapters 3 and 4 are devoted to more practical applications and thus this thesis is naturally divided into theoretical and empirical parts. In this latter part, we analyze statistical properties of end-user sessions by studying two segmentation schemes that help with statistical analysis of network sessions: Segmentation by peak rate (Chapter 3) and by application (Chapter 4). These segmentation schemes will help construct simulation methodology for end-user activities.

In Chapter 3, we refine a stimulating study by Sarvotham et al. (2005) which highlighted the influence of peak transmission rate on network burstiness. From TCP packet headers, we amalgamate packets into sessions where each session is characterized by a 5-tuple $(S, D, R, R^{\vee}, \Gamma)$=(total payload, duration, average transmission rate, peak transmission rate, initiation time). We first introduce a definition of the peak transmission rate. After careful consideration, a new definition of peak rate is required. Unlike Sarvotham et al. (2005) who segmented sessions into two groups labelled alpha and beta, we segment into 10 sessions according to the empirical quantiles of the peak rate variable as a demonstration that the beta group is far from homogeneous. Our more refined segmentation reveals additional structure that is missed by segmentation into two groups. In each segment, we study the dependence structure of (S, D, R) and find that it varies across the groups. Furthermore, within each segment,

session initiation times are well approximated by a Poisson process whereas this property does not hold for the data set taken as a whole. Therefore, we conclude that the peak rate level is important for understanding structure and for constructing accurate simulations of data in the wild. We outline a simple method of simulating network traffic based on our findings.

Finally, in Chapter 4, we summarize extreme value analysis of network applications. Construction of application sessions is a difficult network task, but we use the traditional matching of network applications with well-known ports, and focus on ports 80 (HTTP), 443 (HTTPS), 25 (SMTP) and 1935 (RTMP), which transmit the most bytes. For each port, we look for heavy tails for sizes and durations. An issue here is that of censored observations, which occurs when the network sessions start before or end after the collection interval, but endure within the interval. For such sessions, the traditional POT method and the more recent formal analysis by Dietrich et al. (2002) do not account for the censored portion of the data. Thus, we look at Beirlant and Guillou (2001)'s variant of the Hill estimator for the shape parameter, which helps with the estimation of the distribution of session durations, but not sizes. Furthermore, within each port, session interarrival times are not exponential or independent, but we show that sessions arrive according to a catastrophe process, that is, there is a Poisson process behavior between network disruptions.

CHAPTER 2

SUPERPOSITION OF HETEROGENEOUS TRAFFIC AT LARGE TIME SCALES

2.1 Overview

Collection of data network measurements often uses an algorithm for clustering packets with the same source and destination IP addresses into network sessions (see Section 1.1). Then, a time resolution or granularity is selected or imposed. Typical resolutions are 1, 10 or 100 milliseconds, 1 second, 1 minute, 1 hour, etc. Once a resolution is fixed, the number of bytes or number of packets per unit time can be recorded and *cumulative network loads* over stationary time intervals computed. These cumulative loads have been studied from empirical and theoretical perspectives with the objectives of satisfying performance criterion and offering adequate bandwidth provisioning (van de Meent and Mandjes, 2005) or predicting properties of congestion events (Jin et al., 2007).

Conventional wisdom based on empirical studies claims that a heavily loaded network link subject to aggregation over many users should see Gaussian traffic. This wisdom is considered a network *invariant*. Influential examples based on the Bellcore measurements (Leland et al., 1994) suggest that *horizontal* aggregation, that is, working with a single on/off stream at sufficiently large time scale justifies Gaussian modeling. See also Kurtz (1996) and Willinger et al. (1997).

However, mathematically it is known that with heavy tailed session durations, cumulative load at large time scales can be approximated by either frac-

tional Brownian motion (fBm) or stable Lévy motion, depending on whether the intensity at which sessions are initiated is large or small relative to the size of the duration tails. See Mikosch et al. (2002); Kaj and Taqqu (2008); Taqqu et al. (1997). The stable approximation has not been observed empirically (Guerin et al., 2003) and use of Gaussian cumulative loads has become dominant (Kilpi and Norros, 2002; Sarvotham et al., 2002; Jain and Dovrolis, 2005).

But why should traffic be Gaussian? According to the empirical study van de Meent et al. (2006), in addition to horizontal aggregation, the superposition of independent traffic streams, that is, *vertical* aggregation, can justify a Gaussian model and, in fact, the number of traffic streams need not be large to make cumulative loads approximately Gaussian.

In this chapter we

- study the distribution of the cumulative load in the presence of a finite number of independent traffic streams;
- give an explanation for why Gaussian examples abound in practice but not stable ones;
- answer how much vertical aggregation is needed to justify the use of fBm.

Our findings suggest that cumulative load for aggregate traffic can be approximated by fBm at large time scales provided the initiation intensity of at least one of the traffic components is large. Network traffic in the wild has several distinct constituents and we claim that in practice there is one or more components with dominant large initiation intensities. For example, this should be the case with web traffic using port 80 and this suggests why Gaussian traffic should be pervasive (van de Meent et al., 2006).

Before discussing mathematical details, we illustrate the phenomena of interest with a motivating example of a network trace captured at Cornell University main campus servers during 55 days between November 2, 2009, and January 15, 2010. Cornell's data set is a collection of *netflow* records, where only TCP and UPD traffic is present in the trace. A netflow is a collection of packets with the same source and destination IP addresses, source and destination ports, protocol, ingress interface and IP type of service (Cisco Systems, Inc.). In our data, TCP traffic accounts for nearly 90% of the bytes, and over 80% of the total number of netflows, mostly port 80 (http traffic) netflows. We have taken the part of the trace corresponding to both outgoing and incoming traffic between 1 and 5 p.m. local time, adding up to 220 hours of traffic. The anonymization procedure used on the data obliterated the distinction between outgoing and incoming flows.

We analyze the distribution of $A^{(TCP)}$ and $A^{(UDP)}$, namely the cumulative load generated by TCP and UDP bytes, respectively. For this purpose, we separate the trace into TCP and UDP netflows and for $k = 1, \ldots, 220$ we count

$$A_k^{(TCP)} := \text{total number of TCP bytes captured in the } k\text{th hour},$$
$$A_k^{(UDP)} := \text{total number of UDP bytes captured in the } k\text{th hour}.$$

Due to the dates and times of collection, these counts exhibit both a trend and a daily seasonality. Here we detrend and remove daily seasonality (see e.g. Brockwell and Davis, 1991, Section 1.4), but our conclusions are the same without this massage.

Figure 2.1 shows Gaussian QQ plots for $A^{(TCP)}$ (*left*) and $A^{(UDP)}$ (*right*). A straight line fit is evident for the TCP cumulative input. However, the UDP counterpart shows a significant departure from the straight line. Using the

p-values of the Anderson-Darling two-sided test also shows no evidence against the normality of $A^{(TCP)}$ ($p = 0.1369$), but strong evidence against a Gaussian model for $A^{(UDP)}$ ($p = 9.8 \times 10^{-16}$).

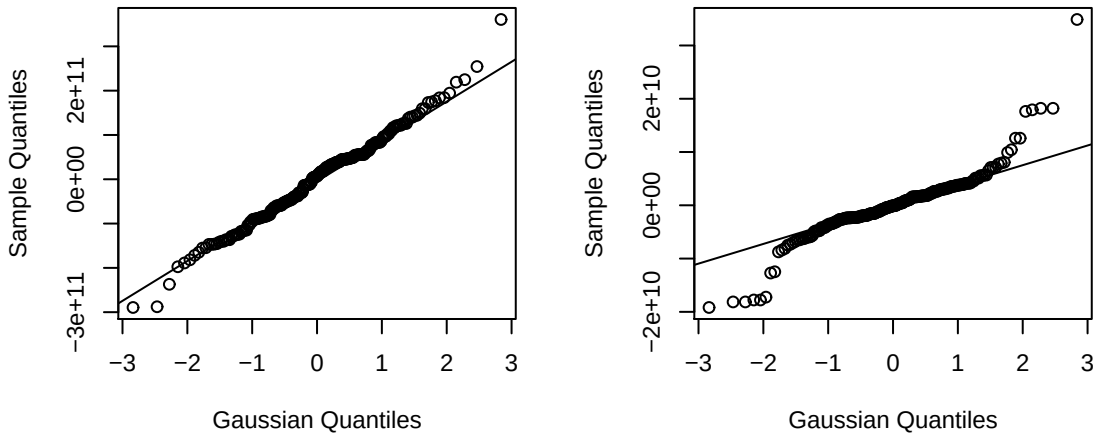

Figure 2.1: Normal QQ plots of cumulative inputs. *Left*: TCP traffic. *Right*: UDP traffic.

We also check whether $A^{(UDP)}$ is a heavy-tailed random variable, in the sense of its distribution tail being regularly varying with tail index α (de Haan and Ferreira, 2006; Resnick, 2007). For instance, Figure 2.2 *left* shows a stable regime in the Hill plot of α (for Hill plots, see e.g. Hill, 1975; de Haan and Resnick, 1998; Resnick, 2007). Additionally, in Figure 2.2 *right* we present the exponential QQ plot of $\ln A^{(UDP)}$ with a straight line fit through the biggest 55 observations. This shows no evidence against approximating the distribution of thresholded values of $A^{(UDP)}$ by a Pareto. (Recall that the logarithm of Pareto random variable is exponential; see Section 1.2.)

If we consider the aggregated cumulative load, $A^{(TCP)} + A^{(UDP)}$, the normal

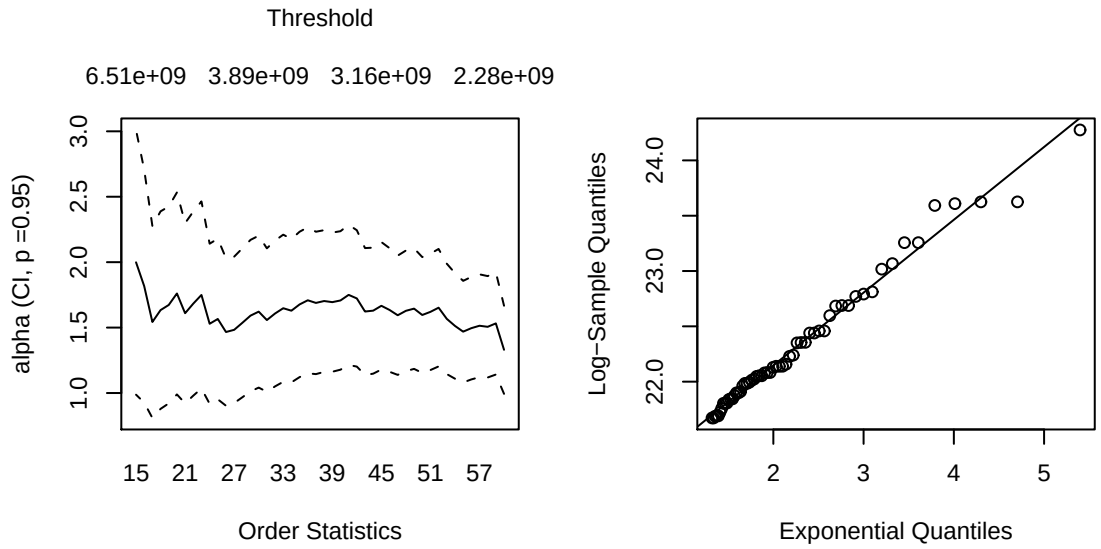

Figure 2.2: Plots for the UDP cumulative input. *Left:* Hill plot of tail index with 95% confidence interval. *Right:* Exponential QQ plot of log-data.

QQ plot in Figure 2.3 exhibits a straight line fit and the Anderson-Darling test $p-$value is 0.2117, showing no evidence to reject normality. Without accounting for centering and scaling, this result is rather counterintuitive due to the nature of the individual tails of $A^{(TCP)}$ and $A^{(UDP)}$.

Our explanation to the above phenomenon starts by modeling the quantity of data in windows of length T in Section 2.2. Analogously to the slow and fast growths of Mikosch et al. (2002), we define two different scenarios for the aggregated traffic. A third scenario is defined similarly to the boundary case considered in Kaj and Taqqu (2008). In Section 2.3 we obtain approximations and provide clarification of the asymptotic behavior at large time scales. We let $T \to \infty$ and see what limits exist for the aggregated cumulative load. In Section 2.4 we study extensions to our model and finally Section 2.5 contains some technical results used to prove our main theorems.

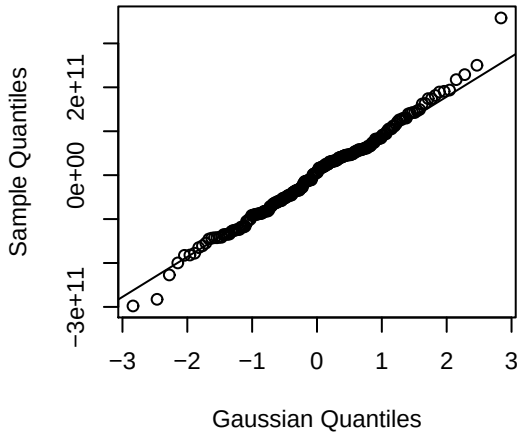

Figure 2.3: Normal QQ plot of the aggregated cumulative input.

2.2 Model description and basic assumptions

We consider a slight modification of the infinite-source Poisson model (Section 1.3). Our network also has an infinite number of nodes. At certain times, a node begins a transmission session at a random rate that is fixed throughout the session. We now suppose network traffic consists of p distinct types which we call *streams*. In practice, such a division of network traffic arises naturally; e.g. traffic can be segmented by application type (web, email, streaming media, file-sharing applications, etc.), by transport protocol (TCP, UDP, IMTP, etc.), and even by users. We suppose the p streams are independent and that each follows an $M/G/\infty$ input model. The overall load is obtained by aggregating over the p streams. Thus, the basic assumptions are as follows:

- Sessions of the jth stream are initiated at homogenous Poisson time points $\{\Gamma_k^{(j)}, -\infty < k < \infty\}$ with arrival intensity $\lambda^{(j)} > 0$. These points are

labeled so that $\Gamma_0^{(j)} < 0 < \Gamma_1^{(j)}$ whence $\{-\Gamma_0^{(j)}, \Gamma_1^{(j)}, (\Gamma_{k+1}^{(j)} - \Gamma_k^{(j)}, k \neq 0)\}$ are iid exponential with parameter $\lambda^{(j)}$. Thus, we have:

$$\sum_k \epsilon_{\Gamma_k^{(j)}} = \mathbb{PRM}(\lambda^{(j)} ds).$$

We assume that these $\mathbb{PRM}$s are independent.

- All the sessions in the network transmit data at positive random rates that are iid with common distribution F_R. Let $\{R_k^{(j)}\}$ be the rate of the kth session of the jth stream. Assume that either $\bar{F}_R \in RV_{-\alpha_R}, 1 < \alpha_R < 2$, or $E[(R_1^{(1)})^2] < \infty$. In either case, define $\mu_R := ER_1^{(1)}$.
- Sessions in the jth stream have positive durations $\{D_k^{(j)}\}$, $j = 1, \ldots, p$, that are iid $F_D^{(j)}$, with $\bar{F}_D^{(j)} \in RV_{-\alpha_D^{(j)}}$, $1 < \alpha_D^{(j)} < 2$, and $\mu_D^{(j)} := ED_1^{(j)}$. In general, not all the $\alpha_D^{(j)}$s are equal.
- We also assume mutually independent durations across streams, and that durations and rates are independent.

There is empirical evidence justifying the choices of $\alpha_D^{(j)}$s and α_R: See e.g. Cunha et al. (1995); Willinger et al. (1995); Leland et al. (1994); Resnick (2003); López-Oliveros and Resnick (2011). For now, we adopt a network-centric approach by assuming the rate of communication entirely depends on the state and speed of the network. Studies supporting this assumption include Shakkottai et al. (2005) and Kortebi et al. (2005).

We will need

$$\lambda = \sum_{j=1}^{p} \lambda^{(j)}, \tag{2.1}$$

$$F_D := \sum_{j=1}^{p} (\lambda^{(j)}/\lambda) F_D^{(j)}, \tag{2.2}$$

and the quantile functions

$$b_D^{(j)}(t) = (1/\bar{F}_D^{(j)})^{\leftarrow}(t) = (F_D^{(j)})^{\leftarrow}(1 - 1/t), \tag{2.3}$$

$$b_D(t) = (1/\bar{F}_D)^{\leftarrow}(t) = F_D^{\leftarrow}(1 - 1/t), \tag{2.4}$$

$$b_R(t) = (1/\bar{F}_R)^{\leftarrow}(t) = F_R^{\leftarrow}(1 - 1/t). \tag{2.5}$$

Notice that F_D is the mixture model of the durations of the p streams, with weights $\lambda^{(j)}/\lambda$, $j = 1, \ldots, p$. In fact, F_D is the distribution of the duration of the sessions of the aggregated stream, and $\lambda^{(j)}/\lambda$ is the proportion of the traffic that consists of sessions from the jth stream. We return to this interpretation later.

Now consider (s, u, r) as a generic Poisson point representing a session that starts at time s, has duration u and rate r. By augmentation, the counting function of the session descriptors $(\Gamma_k^{(j)}, D_k^{(j)}, R_k^{(j)})$ of the jth stream on $\mathbb{R} \times [0, \infty)^2$ is

$$N^{(j)} := \sum_k \epsilon_{(\Gamma_k^{(j)}, D_k^{(j)}, R_k^{(j)})} = \mathbb{PRM}(\lambda^{(j)} ds F_D^{(j)}(du) F_R(dr)), \quad j = 1, \ldots, p. \tag{2.6}$$

By independence, the counting function of the session descriptors of the aggregated stream is

$$\begin{aligned} N := \sum_{j=1}^{p} N^{(j)} &= \mathbb{PRM}\left(\lambda ds \sum_{j=1}^{p} (\lambda^{(j)}/\lambda) F_D^{(j)}(du) F_R(dr)\right) \\ &= \mathbb{PRM}(\lambda ds F_D(du) F_R(dr)). \end{aligned} \tag{2.7}$$

Thus, the mean measures of the $N^{(j)}$ and N are given by

$$EN^{(j)}(ds, du, dr) := \lambda^{(j)} ds F_D^{(j)}(du) F_R(dr), \quad j = 1, \ldots, p,$$

$$EN(ds, du, dr) := \lambda ds F_D(du) F_R(dr).$$

In addition, let

$$L_t(s, u) = |[0, t] \cap [s, s + u]| = \int_0^t 1_{[s, s+u]}(y) dy = \int_0^u 1_{[0,t]}(y + s) dy, \tag{2.8}$$

be the length of the subinterval of $[0,t]$ during which the session (s,u,r) transmits data. In Lemma 2.5.4, we summarize several required properties of $L_t(s,u)$.

For each j, define

$$\begin{aligned} A^{(j)}(t) &:= \text{cumulative input in } [0,t] \text{ from the } j\text{th stream} \\ &= \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} rL_t(s,u)N^{(j)}(ds,du,dr), \end{aligned} \tag{2.9}$$

and similarly

$$\begin{aligned} A(t) &:= \text{cumulative input in } [0,t] \text{ from the aggregated stream} \\ &= \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} rL_t(s,u)N(ds,du,dr). \end{aligned} \tag{2.10}$$

(Kaj and Taqqu (2008) showed that these integrals are well defined using Campbell's theorem (Kingman, 1993, Section 3.2).) Also,

$$EA^{(j)}(t) = \lambda^{(j)}\mu_D^{(j)}\mu_R t, \quad EA(t) = \sum_{j=1}^{p}\lambda^{(j)}\mu_D^{(j)}\mu_R t = \lambda\mu_D\mu_R t,$$

where

$$\mu_D := \sum_{j=1}^{p}(\lambda^{(j)}/\lambda)\mu_D^{(j)} \tag{2.11}$$

is the mean of the mixture model of the durations of the different streams.

Observe that we can write the cumulative inputs as linear drift plus compensated random Poisson fluctuation as follows:

$$A^{(j)}(t) := \lambda^{(j)}\mu_D^{(j)}\mu_R t + \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} rL_t(s,u)\mathring{N}^{(j)}(ds,du,dr), \tag{2.12}$$

$$A(t) := \lambda\mu_D\mu_R t + \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} rL_t(s,u)\mathring{N}(ds,du,dr). \tag{2.13}$$

After scaling time by T, we think of $A_T^{(j)} := (A^{(j)}(Tt), t>0), j=1,\ldots,p$, and $A_T := (A(Tt), t>0)$ for large T, as the cumulative inputs on large time scales.

Thus, we consider a family of models indexed by the time scale parameter T and from now on we let the arrival intensities depend on T so that $\lambda^{(j)} := \lambda^{(j)}(T)$. If necessary, we let $\lambda_j(T) \to \infty$ as $T \to \infty$ (see (2.15)). Dependence of the arrival intensities on T means λ, F_D, $b_D^{(j)}$ and b_D as defined in (2.1)-(2.4) depend on T as well; however, notice that the tail indices of the distribution of the duration, namely $\alpha_D^{(j)}$, remain independent of T. In practice, the fact that we focus on the stream at a particular time period, say $[0, Tt]$, does not affect the tail index of the distribution of the sessions duration, which is in accordance with our assumptions. For convenience, we often suppress the subscript T.

Fix j, $1 \leq j \leq p$ and in the Tth model, let $A_{cs}^{(j)}(t)$ be the centered and scaled cumulative input of the jth stream in $[0, Tt]$, that is

$$A_{cs}^{(j)}(t) := \frac{A^{(j)}(Tt) - \lambda^{(j)}\mu_D^{(j)}\mu_R Tt}{a^{(j)}(T)}, \tag{2.14}$$

for a suitable $a_j(T)$ to be made precise below. Assuming $\lim_{T\to\infty} \lambda^{(j)} T \bar{F}_D^{(j)}(T)$ exists, the asymptotic behavior of $A_{cs}^{(j)}(t)$ as $T \to \infty$, depends on whether the arrival rate is large, moderate, or small, relative to the tail of the duration.

Theorem 2.2.1. *(Mikosch et al., 2002; Kaj and Taqqu, 2008).*

For any $1 \leq j \leq p$, consider the following three growth regimes of the arrival rate:

$$\lim_{T\to\infty} \lambda^{(j)} T \bar{F}_D^{(j)}(T) = \begin{cases} \infty, & \textit{fast-growth}. \\ c_j^{\alpha_D^{(j)}-1}, & \textit{moderate-growth}, \\ 0, & \textit{slow-growth}, \end{cases} \tag{2.15}$$

where $c_j \in (0, \infty)$. (The form of the moderate-growth limit facilitates a simple expression of the corresponding limit process.) Assume that either $\bar{F}_R \in RV_{-\alpha_R}, \alpha_R > \alpha_D^{(j)}$ or $E[(R_1^{(1)})^2] < \infty$. (If $\alpha_R \leq \alpha_D^{(j)}$, the limit process is the same for all three growth

regimes and the distinction among the growth regimes is irrelevant (Kaj and Taqqu, 2008, Theorem 4).)

(a) Under fast-growth, we distinguish two subcases:

(i) If $E[(R_1^{(1)})^2] < \infty$,

$$A_{cs}^{(j)}(\cdot) \xrightarrow{fidi} E[(R_1^{(1)})^2]^{1/2} \sigma^{(j)}_{B_{H^{(j)}}(1)} B_{H^{(j)}}(\cdot), \quad T \to \infty,$$

where

$$a^{(j)}(T) = [\lambda^{(j)} T^3 \bar{F}_D^{(j)}(T)]^{1/2},$$

$$\sigma^{(j)}_{B_{H^{(j)}}(1)} = \frac{2}{(\alpha_D^{(j)} - 1)(2 - \alpha_D^{(j)})(3 - \alpha_D^{(j)})},$$

and $B_{H^{(j)}}$ *is a fractional Brownian motion with Hurst exponent*

$$H^{(j)} = (3 - \alpha_D^{(j)})/2 \in (1/2, 1).$$

(ii) If $\bar{F}_R \in RV_{-\alpha_R}, 1 < \alpha_D^{(j)} < \alpha_R < 2$, *then*

$$A_{cs}^{(j)}(\cdot) \xrightarrow{fidi} Z_{\alpha_D^{(j)}, \alpha_R}(\cdot), \quad T \to \infty,$$

where

$$a^{(j)}(T) = T b_R(\lambda T \bar{F}_D^{(j)}(T)),$$

$$Z_{\alpha_D^{(j)}, \alpha_R}(t) = \int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} r L_t(s,u) \mathring{N}^{\infty}_{\alpha_D^{(j)}, \alpha_R}(ds, du)$$

$$\stackrel{d}{=} \left(\left(-\cos \frac{\pi \alpha_D^{(j)}}{2} \right) \frac{2\Gamma(2 - \alpha_D^{(j)})}{\alpha_D^{(j)}(\alpha_D^{(j)} - 1)} \right)^{1/\alpha_D^{(j)}} \int_{-\infty}^{\infty} \int_0^{\infty} L_t(s,u) M_{\alpha_R, m}(ds, du),$$

and $M_{\alpha_R, m}(ds, du)$ *is a* α_R*-stable random measure with control measure*

$$m(ds, du) = ds \cdot \alpha_D^{(j)} u^{-(\alpha_D^{(j)} + 1)} du.$$

Thus, the process $Z_{\alpha_D^{(j)}, \alpha_R}(t)$ *is* α_R*-stable and* $H^{(j)}$*-similar with*

$$H^{(j)} = (\alpha_R + 1 - \alpha_D^{(j)})/\alpha_R \in (1/\alpha_R, 1).$$

(b) Under moderate-growth

$$A_{cs}^{(j)}(\cdot) \xrightarrow{fidi} c_j Y_{\alpha_D^{(j)}}(\cdot/c_j), \quad T \to \infty,$$

where

$$a^{(j)}(T) = T,$$

and

$$Y_{\alpha_D^{(j)}}(t) = \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} rL_t(s,u)\mathring{N}^{\infty}_{\alpha_D^{(j)},F_R}(ds,du,dr).$$

(c) Under slow-growth

$$A_{cs}^{(j)}(\cdot) \xrightarrow{fidi} E[(R_1^{(1)})^{\alpha_D^{(j)}}]^{1/\alpha_D^{(j)}} \Lambda_{\alpha_D^{(j)}}(\cdot), \quad T \to \infty,$$

where

$$a^{(j)}(T) = b_D^{(j)}(\lambda^{(j)}T),$$

$\Lambda_{\alpha_D^{(j)}}$ *is an* $\alpha_D^{(j)}$*-stable Lévy motion totally skewed to the right, which we can write as*

$$E[(R_1^{(1)})^{\alpha_D^{(j)}}]^{1/\alpha_D^{(j)}} \Lambda_{\alpha_D^{(j)}}(t) = \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} ur1_{\{0<s<t\}}\mathring{N}^{\infty}_{\alpha_D^{(j)},F_R}(ds,du,dr)$$

$$\stackrel{d}{=} \left(\left(-\cos\frac{\pi\alpha_D^{(j)}}{2}\right)\frac{2\Gamma(2-\alpha_D^{(j)})}{\alpha_D^{(j)}(\alpha_D^{(j)}-1)}\right)^{1/\alpha_D^{(j)}} \int_{-\infty}^{\infty}\int_0^{\infty} r1_{\{0<s<t\}}M_{\alpha_D^{(j)},m}(ds,dr),$$

and $M_{\alpha_D^{(j)},m}(ds,dr)$ *is an* $\alpha_D^{(j)}$*-stable random measure with control measure*

$$m(ds,dr) = dsF_R(dr).$$

Real network traffic consists of several distinct types and in this paper we are interested in the centered and scaled cumulative input of the superimposed streams in $[0,Tt]$, namely

$$A_{cs}(t) := \frac{A(Tt) - \lambda\mu_D\mu_R Tt}{a(T)}, \tag{2.16}$$

for a suitable $a(T)$. In order to study the limit distribution of $A_{cs}(t)$ as $T \to \infty$, let $\mathcal{F}$, $\mathcal{M}$, $\mathcal{S}$ be the subsets of indices of streams whose arrival intensities behave under the fast-, moderate-, and slow-growth regimes, respectively.

Assuming that all indices belong to one of these three classes, consider the following scenarios.

Scenario $\mathcal{F}$: There is at least one stream whose arrival intensity satisfies fast-growth; i.e. $\mathcal{F} \neq \emptyset$. In this case, the aggregated stream's arrival intensity also satisfies fast-growth:

$$\lambda T \bar{F}_D(T) \geq \sum_{j \in \mathcal{F}} \lambda^{(j)} T \bar{F}_D^{(j)}(T) \to \infty, \quad T \to \infty. \tag{2.17}$$

Scenario $\mathcal{M}$: No stream's arrival intensity satisfies fast-growth, but at least one stream satisfies moderate-growth; i.e. $\mathcal{F} = \emptyset$ and $\mathcal{M} \neq \emptyset$. Then, the aggregated stream's arrival intensity satisfies moderate growth, since

$$\lambda T \bar{F}_D(T) \to c^{\alpha_D - 1}, \quad T \to \infty, \tag{2.18}$$

where

$$\alpha_D := \bigwedge_{j=1}^{p} \alpha_D^{(j)} \tag{2.19}$$

and

$$c = \left(\sum_{j \in \mathcal{M}} c_j^{\alpha_D^{(j)} - 1} \right)^{1/(\alpha_D - 1)}. \tag{2.20}$$

Scenario $\mathcal{S}$: All the stream's arrival intensities satisfy slow growth, that is $\mathcal{S} = \{1, \ldots, p\}$. In this case, the aggregated stream's arrival intensity also satisfies slow-growth:

$$\lambda T \bar{F}_D(T) = \sum_{j \in \mathcal{S}} \lambda^{(j)} T \bar{F}_D^{(j)}(T) \to 0, \quad T \to \infty. \tag{2.21}$$

The different growth regimes in Theorem 2.2.1 are specified by the arrival intensity $\lambda^{(j)}$, and the distribution $F_D^{(j)}$ of the duration of the sessions of the jth stream. While $\lambda^{(j)} = \lambda^{(j)}(T) \to \infty$ as $T \to \infty$, $F_D^{(j)}$ does not vary with T. However, the growth regimes described in Scenarios $\mathcal{F}$, $\mathcal{M}$ and $\mathcal{S}$ are given in terms of the arrival rate λ, and the distribution F_D of the duration of the sessions of the aggregated stream and here both λ and F_D vary with T, as seen in (2.2). Therefore, we cannot directly apply Theorem 2.2.1 for the aggregated stream when

$$\lambda^{(j)}/\lambda = \text{proportion of the sessions that belong to the } j\text{th stream}, j = 1, \ldots, p, \tag{2.22}$$

are functions of T. Nevertheless, in the special case that these proportions are constant, F_D does not vary with T, and a direct application of Theorem 2.2.1 yields the following result.

Corollary 2.2.2. *Suppose that for all T (or at least for T large enough), the proportions $\lambda^{(j)}/\lambda$ remain constant, $j = 1, \ldots, p$, so that*

$$\bar{F}_D = \sum_{j=1}^{p} (\lambda^{(j)}/\lambda) \bar{F}_D^{(j)} \in RV_{-\alpha_D},$$

where α_D is given in (2.19)*. Let the Scenarios $\mathcal{F}$, $\mathcal{M}$ and $\mathcal{S}$ take the place of the fast-, moderate- and slow-growth regimes.*

If $\bar{F}_R \in RV_{-\alpha_R}, \alpha_R > \alpha_D$ or $E[(R_1^{(1)})^2] < \infty$, then Theorem 2.2.1 holds for $A_{cs}(\cdot)$, where $\alpha_D^{(j)}$, $F_D^{(j)}$ and c_j are replaced by α_D, F_D and the constant c in (2.20)*, respectively.*

If $\bar{F}_R \in RV_{-\alpha_R}, \alpha_R \leq \alpha_D$, the distinction among Scenarios $\mathcal{F}$, $\mathcal{M}$ and $\mathcal{S}$ is irrelevant, and limit results are discussed in Section 2.4.

Implications of Corollary 2.2.2. This result provides a partial answer to the question of how much aggregation is required for traffic to be Gaussian at large time scales: Suppose that at least one traffic stream falls in the fast-growth regime, thus generating a cumulative input that can be approximated by fractional Brownian motion. When applicable, Corollary 2.2.2 implies that the superimposed traffic load can also be approximated by fractional Brownian motion.

In the case that the traffic also contains streams that satisfy the slow-growth regime, Corollary 2.2.2 is somewhat counterintuitive due to the nature of the distribution tails of the two limit processes. Although these slow-growth streams produce cumulative inputs that are approximately stable Lévy-motion when considered individually, with the inclusion of one single stream that behaves under the fast-growth regime, the cumulative aggregated input is approximately Gaussian.

Moreover, a sufficient condition for the fast-growth regime of Scenario $\mathcal{F}$ is that a single stream, say the jth one, satisfies fast-growth, even if all the other streams' arrival intensities do not follow a growth regime at all. In this sense, Scenario $\mathcal{F}$ is a robust assumption. We will see that as long as one $\alpha_D^{(j)} < \alpha_R$, the limit result of Corollary 2.2.2 is still valid.

In real networks, there are arguably streams with large initiation rates. For instance, the arrival rates of http traffic must be large, since there are a large number of users constantly accessing websites and this translates into Scenario $\mathcal{F}$. Furthermore, even though some studies report or assume session transmission rates have infinite variance, the assumption $E[(R_1^{(1)})^2] < \infty$ may be justified by rate constraint mechanisms required for congestion control. Although

assumptions always deserve rigorous scrutiny, Corollary 2.2.2 provides a compelling explanation for the data example in Section 3.1.

We now address more general assumptions which allow the conclusions of Corollary 2.2.2 to hold. While the assumption of constant proportions $\lambda^{(j)}/\lambda$ may sometimes be reasonable, in general the proportions of sessions corresponding to the p independent streams are not constant over time. We may have that $\lim \lambda^{(j)}/\lambda$ exists or, more generally, that $\lambda^{(j)}/\lambda \in (a, b) \subset (0, 1)$ varies with no limit whatsoever. Extending the conclusions of Corollary 2.2.2 to under weaker assumptions is the focus of the next section.

2.3 Behavior of cumulative load of aggregated streams

We prove that the conclusion of Corollary 2.2.2 is still valid even when the proportion of the sessions corresponding to the p independent streams is not constant. Here is the result:

Theorem 2.3.1. *Assume that*

$$\liminf_{T\to\infty} \bigvee_{j:\alpha_D^{(j)}=\alpha_D} \lambda^{(j)}/\lambda > 0. \tag{2.23}$$

Then, the conclusions of Corollary 2.2.2 regarding the limit distribution of the cumulative input of the aggregated stream $A_{cs}(\cdot)$ are still valid.

Condition 2.23 implies that there exists $d > 0$ such that for all T sufficiently large, there is at least one $k = k(T)$ such that $\alpha_D^{(k)} = \alpha_D$ and $\lambda^{(k)}/\lambda > d$. Roughly speaking, this means that the proportion of the traffic with the heaviest-tailed duration always remains greater than a positive quantity.

All the limits in Theorem 2.3.1 follow from the convergence of the characteristic function of the finite-dimensional distributions (*fidi chf*) of the processes. Thus, let $m \geq 1$ represent the dimension, $0 \leq t_1, \ldots, t_m$ the times, and $z_1, \ldots, z_m$ arbitrary real numbers; we need

$$g(s,u,r) = \exp\left\{ i \sum_{j=1}^{m} z_j r L_{t_j}(s,u) \right\} - 1 - i \sum_{j=1}^{m} z_j r L_{t_j}(s,u),$$

as defined in Proposition 2.5.6.

From the second integral in (2.8), we can compute the partial derivative of $L_t(s,u)$ with respect to u, which yields

$$g_u(s-u,u,r) := \frac{\partial}{\partial u} g\Big|_{(s-u,u,r)} = \\ i \left(\exp\left\{ i \sum_{j=1}^{m} z_j r L_{t_j}(s-u,u) \right\} - 1 \right) \sum_{k=1}^{m} z_k r 1_{[0,t_k]}(s), \tag{2.24}$$

where g_u is the partial derivative of $g(s,u,r)$ with respect to u. Moreover, putting together (2.50), (2.62), (2.64), the bounds in Lemmas 2.5.4 and 2.5.7, we get

$$\left| \exp\left\{ i \sum_{j=1}^{m} z_j r L_{t_j}(s-u,u) \right\} - 1 \right| \leq 2 \sum_{j=1}^{m} |z_j|^{\zeta} (t_j \wedge u)^{\zeta} r^{\zeta}, \quad 0 \leq \zeta \leq 1. \tag{2.25}$$

We will use three more relations in the proof of Theorem 2.3.1: For $0 < \eta < 1$, there exists a number $T_0 = T_0(\eta) > 0$ such that for $T \geq T_0$ and $b_D(\lambda T) \geq T_0$,

$$2u^{-\alpha_D} \left\{ u^{-\eta} \vee u^{\eta} \right\} \geq \begin{cases} \bar{F}_D(Tu)/\bar{F}_D(T), & u \geq T_0/T, \\ \bar{F}_D(b_D(\lambda T)u)/\bar{F}_D(b_D(\lambda T)), & u \geq T_0/b(\lambda T), \end{cases} \tag{2.26}$$

$$\frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} \leq \mu_D T^{\alpha_D - 1 + \eta} u^{-1}, \quad u < T_0/T, \tag{2.27}$$

and

$$\frac{\bar{F}_D(b_D(\lambda T)u)}{\bar{F}_D(b_D(\lambda T))} \leq \mu_D b_D(\lambda T)^{\alpha_D - 1 + \eta} u^{-1}, \quad u < T_0/b_D(\lambda T). \tag{2.28}$$

We can readily derive (2.26) from Lemma 2.5.3. Both (2.27) and (2.28) follow from Markov's Inequality and, for example, Resnick (1987, Proposition 0.8) or Bingham et al. (1987, Proposition 1.3.6).

Proof of Theorem 2.3.1. First, we will prove parts (b) and (c). For both parts, set $0 < \eta < \alpha_D - 1$ and $0 < \zeta < 1$ such that

$$\alpha_D + \eta < 1 + \zeta < \begin{cases} \alpha_R, & \text{if } \bar{F}_R \in RV_{-\alpha_R}, 1 < \alpha_R < 2, \\ 2, & \text{if } E[(R_1^{(1)})^2] < \infty. \end{cases}$$

Part (b). Under Scenario $\mathcal{M}$, use $a(T) = T$ and apply Proposition 2.5.6, yielding

$$\ln E \exp\left\{ i \sum_{j=1}^{m} z_m A_{cs}(t_j) \right\} = \int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} g_u(s-u,u,r) \lambda T \bar{F}_D(T) \frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} ds du F_R(dr),$$

and if we can take the limit inside the integral as $T \to \infty$, performing afterwards an integration by parts in u gives

$$\begin{aligned} &\to \int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} g_u(s-u,u,r) c^{\alpha_D - 1} u^{-\alpha_D} ds du F_R(dr) \\ &= c^{\alpha_D - 1} \int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} g(s,u,r) E N^{\infty}_{\alpha_D^{(j)}, F_R}(ds, du, dr), \end{aligned} \tag{2.29}$$

which is the log fidi chf of $cY_{\alpha_D}(\cdot/c)$. Thus, it suffices to justify taking the limit inside the integral.

First observe there exists a number $T_0 > 0$ such that for $T \geq T_0$,

$$\lambda T \bar{F}_D(T) \leq c^{\alpha_D - 1} + \eta, \tag{2.30}$$

by the moderate-growth assumption. Together with (2.25)-(2.27) and a possibly larger T_0, the above implies that the integrand in the left side of (2.29) is bounded

in $\{u \geq T_0/T\}$ by

$$B_{\mathcal{M},(>)}(s,u,r) := 4\left(c^{\alpha_D-1}+\eta\right)u^{-\alpha_D}(u^{-\eta}\vee u^{\eta})\sum_{j=1}^{m}\sum_{k=1}^{m}|z_j|^{\zeta}|z_k|(t_j\wedge u)^{\zeta}r^{1+\zeta}1_{[0,t_k]}(s),$$

and bounded in $\{u < T_0/T\}$ by

$$B_{\mathcal{M},(<)}(s,u,r) := \\ 2\left(c^{\alpha_D-1}+\eta\right)T_0^{\alpha_D-1+\eta}\mu_D\sum_{j=1}^{m}\sum_{k=1}^{m}|z_j|^{\zeta}|z_k|u^{\zeta-\alpha_D-\eta}r^{1+\zeta}1_{[0,t_k]}(s)1_{(0,1)}(u),$$

whenever $T \geq T_0$. Here we used the bound

$$u^{\zeta} \leq (T_0/T)^{\alpha_D-1+\eta}u^{1+\zeta-\alpha_D-\eta}, \quad 0 < u < T_0/T. \tag{2.31}$$

Therefore, (2.29) follows by the dominated convergence theorem, since for all $T \geq T_0$

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty}B_{\mathcal{M},(>)}(s,u,r)dsduF_R(dr)\\ &\qquad\leq 4\left(c^{\alpha_D-1}+\eta\right)E[(R_1^{(1)})^{1+\zeta}]\times\\ &\qquad\quad\sum_{k=1}^{m}\sum_{j=1}^{m}|z_j|^{\zeta}|z_k|t_k\left\{\int_0^1 u^{\zeta-\alpha_D-\eta}du + t_j^{\zeta}\int_1^{\infty}u^{-\alpha_D+\eta}du\right\},\end{aligned}$$

and

$$\begin{aligned}&\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty}B_{\mathcal{M},(<)}(s,u,r)dsduF_R(dr)\\ &\leq 2\left(c^{\alpha_D-1}+\eta\right)T_0^{\alpha_D-1+\eta}\mu_D\sum_{j=1}^{m}\sum_{k=1}^{m}|z_j|^{\zeta}|z_k|t_kE[(R_1^{(1)})^{1+\zeta}]\int_0^1 u^{\zeta-\alpha_D-\eta}du,\end{aligned}$$

which are both finite by our choice of η and ζ.

Part (c). Under Scenario $\mathcal{S}$, $a(T) = b_D(\lambda T)$, so we use Lemma 2.5.1 and Mikosch et al. (2002, Lemma 1) to get

$$\lim_{T\to\infty} T/a(T) = \infty.$$

Thus, it follows from the definition of $L_t(s,u)$ in (2.8) that

$$\lim_{T\to\infty} L_{tT/a(T)}(sT/a(T)-u,u) = u1_{[0,t]}(s).$$

Now, apply Proposition 2.5.6, perform the change of variables $r \mapsto ra(T)/T$, $u \mapsto uT/a(T)$, and use the scaling property in Lemma 2.5.4 to get

$$\begin{aligned}
&\ln E\exp\left\{i\sum_{j=1}^m z_m A_{cs}(t_j)\right\} \\
&= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} g_u(s-ua(T)/T, ua(T)/T, rT/a(T))\lambda a(T)\bar{F}_D(a(T)u)dsduF_R(dr) \\
&= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} i\left(\exp\left\{\sum_{j=1}^m z_j r L_{t_jT/a(T)}(sT/a(T)-u,u)\right\}-1\right) \\
&\qquad\qquad \times\sum_{k=1}^m z_k r 1_{(0,t_k)}(s)\lambda T\bar{F}_D(a(T)u)dsduF_R(dr),
\end{aligned}$$

and assuming we can take the limit inside the integral, the limit as $T\to\infty$ is

$$\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} i\left(\exp\left\{\sum_{j=1}^m z_j u r 1_{(0,t_j)}(s)\right\}-1\right)\sum_{k=1}^m z_k r 1_{(0,t_k)}(s)u^{-\alpha_D}dsduF_R(dr), \tag{2.32}$$

which is the log fidi chf of $E[(R_1^{(1)})^{\alpha_D}]^{1/\alpha_D}\Lambda_{a_D}(\cdot)$. Therefore, we must justify passing the limit inside the integral. This is done as follows.

First, by Lemma 2.5.2, there exists a number $T_0>0$ such that for $T\ge T_0$,

$$\lambda T\bar{F}_D(a(T)) \le 2. \tag{2.33}$$

Hence, by taking a possibly larger T_0, (2.25), (2.26) and (2.28) imply that the integrand in the left side of (2.32) is bounded in $\{u\ge T_0/a(T)\}$ by

$$B_{\mathcal{S},(>)}(s,u,r) := 8u^{-\alpha_D}(u^{-\eta}\vee u^{\eta})\sum_{j=1}^m\sum_{k=1}^m |z_j|^{\zeta}|z_k|(t_j\wedge u)^{\zeta}r^{1+\zeta}1_{[0,t_k]}(s),$$

and bounded in $\{u < T_0/a(T)\}$ by

$$B_{\mathcal{S},(<)}(s,u,r) := 4T_0^{\alpha_D-1+\eta}\mu_D \sum_{j=1}^{m}\sum_{k=1}^{m} |z_j|^{\zeta}|z_k| u^{\zeta-\alpha_D-\eta} r^{1+\zeta} 1_{[0,t_k]}(s) 1_{(0,1)}(u),$$

whenever $T \geq T_0$ and $a(T) \geq T_0$, using

$$u^{\zeta} \leq (T_0/a(T))^{\alpha_D-1+\eta} u^{1+\zeta-\alpha_D-\eta}, \quad 0 < u < T_0/a(T).$$

Therefore, (2.32) follows exactly as in part (b) from the dominated convergence theorem.

Part (a). Under Scenario $\mathcal{F}$ and $E[(R_1^{(1)})^2] < \infty$, set $a(T) = [\lambda T^3 \bar{F}_D(T)]^{1/2}$. Use Proposition 2.5.6 and the change of variables $r \mapsto ra(T)/T$, to write

$$\ln E \exp\left\{ i \sum_{j=1}^{m} z_m A_{cs}(t_j) \right\} =$$
$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} g_u(s-u,u,rT/a(T))(a(T)/T)^2 \frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} ds du F_R(dr), \qquad (2.34)$$

where

$$(a(T)/T)^2 = \lambda T \bar{F}_D(T) \to \infty, \quad T \to \infty,$$

by the fast-growth assumption.

By (2.24), as $T \to \infty$

$$g_u(s-u,u,rT/a(T))(a(T)/T)^2 =$$
$$i\left(i \sum_{j=1}^{m} z_j r \frac{T}{a(T)} L_{t_j}(s-u,u) + o\left(\frac{T}{a(T)}\right) \right) \sum_{k=1} z_k r 1_{[0,t_k]}(s) \frac{a(T)}{T}.$$

Hence, assuming we can pass the limit inside the integral, we use Lemma 2.5.4 (iii) to write

$$
\begin{aligned}
&\lim_{T\to\infty} \ln E\exp\{i\sum_{j=1}^{m} z_m A_{cs}(t_j)\} \\
&= -\sum_{j=1}^{m}\sum_{k=1}^{m} z_j z_k \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} r^2 L_{t_j}(s-u,u)1_{[0,t_k]}(s)u^{-\alpha_D}dsduF_R(dr) \\
&= -E[(R_1^{(1)})^2]\left(\frac{1}{(\alpha_D-1)(2-\alpha_D)(3-\alpha_D)}\right)\times \\
&\left\{\sum_{j=1}^{m}\sum_{k=1}^{j} z_j z_k t_k^{3-\alpha_D} + \sum_{j=1}^{m}\sum_{k=j+1}^{m} z_j z_k \left(t_k^{3-\alpha_D} - (t_k-t_j)^{3-\alpha_D}\right)\right\} \\
&= -\frac{1}{2}E[(R_1^{(1)})^2]\sigma^2_{B_H(1)}\sum_{j=1}^{m}\sum_{k=1}^{m} z_j z_k \frac{1}{2}\left\{|t_j|^{2H}+|t_k|^{2H}-|t_j-t_k|^{2H}\right\},
\end{aligned}
$$

where the last line follows by rearranging of the terms in the sum, $\sigma^2_{B_H(1)}$ is given in (2.56) and $H=(3-\alpha_D)/2$. It remains to prove that we can take the limit inside the integral.

Let $0<\eta<\alpha_D-1$. We use (2.25)-(2.27) with $\zeta=1$ and a possibly larger T_0, which imply that the integrand in (2.34) is bounded in $\{u\geq T_0/T\}$ by

$$
B_{\mathcal{F},(>)} := 4u^{-\alpha_D}(u^{-\eta}\vee u^{\eta})\sum_{j=1}^{m}\sum_{k=1}^{m}|z_j z_k|(t_j\wedge u)r^2 1_{[0,t_k]}(s),
$$

and bounded in $\{u<T_0/T\}$ by

$$
B_{\mathcal{F},(<)} := 2T_0^{\alpha_D-1+\eta}\mu_D\sum_{j=1}^{m}\sum_{k=1}^{m}|z_j z_k|u^{1-\alpha_D-\eta}r^2 1_{[0,t_k]}(s)1_{(0,1)}(u),
$$

whenever $T\geq T_0$. Here we used

$$
u\leq (T_0/T)^{\alpha_D-1+\eta}u^{2-\alpha_D-\eta},\quad 0<u<T_0/T.
$$

The result now follows by the dominated convergence theorem, since

$$
\begin{aligned}
&\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} B_{\mathcal{F},(>)}(s,u,r)dsduF_R(dr) \\
&\leq 4E[(R_1^{(1)})^2]\sum_{j=1}^{m}\sum_{k=1}^{m}|z_j z_k|t_k\left\{\int_0^1 u^{1-\alpha_D-\eta}du + t_j E R_1^{(1)}\int_1^{\infty} u^{-\alpha_D+\eta}du\right\},
\end{aligned}
$$

and

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} B_{\mathcal{F},(<)}(s,u,r)dsduF_R(dr)$$
$$\leq 2T_0^{\alpha_D-1+\eta}\mu_D\sum_{j=1}^{m}\sum_{k=1}^{m}|z_jz_k|t_kE[(R_1^{(1)})^2]\int_0^1 u^{1-\alpha_D-\eta}du,$$

and both bounds are finite by our choice of η.

Finally, still under Scenario $\mathcal{F}$, assume $\bar{F}_R \in RV_{-\alpha_R}, 1 < \alpha_R < 2$. Set $a(T) = Tb_R(\lambda T\bar{F}_D(T))$. By Proposition 2.5.6, an integration by parts in u and the change of variables $s \mapsto s+u$:

$$\begin{aligned}
&\ln E\exp\left\{i\sum_{j=1}^{m} z_m A_{cs}(t_j)\right\}\\
&= \lambda T\bar{F}_D(T)\bar{F}_R(b_R(\lambda T\bar{F}_D(T)))\times\\
&\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} g(s,u,r)ds\frac{F_D(Tdu)}{\bar{F}_D(T)}\frac{F_R(b_R(\lambda T\bar{F}_D(T))dr)}{\bar{F}_R(b_R(\lambda T\bar{F}_D(T)))}\\
&= \lambda T\bar{F}_D(T)\bar{F}_R(b_R(\lambda T\bar{F}_D(T)))\left\{I_{(u>\epsilon,r>\epsilon)}+I_{(u<\epsilon,r>\epsilon)}+I_{(r<\epsilon)}\right\}, \qquad (2.35)
\end{aligned}$$

where $\epsilon > 0$ and we split the integral into three parts according to the domains of integration $\{u > \epsilon, r > \epsilon\}$, $\{u < \epsilon, r > \epsilon\}$ and $\{r < \epsilon\}$, respectively. To establish the limit result, we will take $\lim_{\epsilon\to 0}\lim_{T\to\infty}$ on both sides of (2.35).

Fix $\epsilon > 0$ and start with the first integral. Let

$$\begin{aligned}
\nu_T(du,dr) &:= \left(u\frac{F_D(Tdu)}{\bar{F}_D(T)}\right)\left(r\frac{F_R(b_R(\lambda T\bar{F}_D(T))dr)}{\bar{F}_R(b_R(\lambda T\bar{F}_D(T)))}\right),\\
\nu(du,dr) &:= \alpha_D u^{-\alpha_D}du\,\alpha_R r^{-\alpha_R}dr,\\
G(u,r) &:= \frac{1}{ur}\int_{-\infty}^{\infty} g(s,u,r)ds,
\end{aligned}$$

which allows writing

$$I_{(u>\epsilon,r>\epsilon)} = \int_{\epsilon}^{\infty}\int_{\epsilon}^{\infty} G(u,r)\nu_T(du,dr).$$

The fast-growth regime, regular variation of $\bar{F}_R$, (2.49) and Billingsley (1999, Theorem 2.8) imply $\nu_T \xrightarrow{v} \nu$ as $T \to \infty$. Moreover, $G(u,r)$ is jointly continuous and it follows from Lemmas 2.5.4 and 2.5.7 that $|G(u,r)| \leq d_0 \sum_{j=1}^m |z_j| t_j < \infty$, where d_0 is a positive constant. Therefore:

$$\begin{aligned}\lim_{T\to\infty} I_{(u>\epsilon, r>\epsilon)} &= \int_\epsilon^\infty \int_\epsilon^\infty G(u,r)\nu(du,dr) \\ &= \int_{-\infty}^\infty \int_\epsilon^\infty \int_\epsilon^\infty g(s,u,r)ds \cdot \alpha_D u^{-(\alpha_D+1)}du \cdot \alpha_R r^{-(\alpha_R+1)}dr \quad (2.36)\end{aligned}$$

Now let $0 \leq \zeta, \eta \leq 1$ such that $\alpha_D + \eta < 1 + \zeta < \alpha_R$. By Lemmas 2.5.4 and 2.5.7, there exists $d_\zeta > 0$ such that

$$\left|I_{(u<\epsilon, r>\epsilon)}\right| \leq d_\zeta \sum_{j=1}^m |z_j| t_j \int_\epsilon^\infty r^{1+\zeta} \frac{F_R(b_R(\lambda T \bar{F}_D(T))dr)}{\bar{F}_R(b_R(\lambda T \bar{F}_D(T)))} \int_0^\epsilon u^{1+\zeta} \frac{F_D(Tdu)}{\bar{F}_D(T)}.$$

Furthermore, by fast-growth and regular variation of $\bar{F}_R$

$$\int_\epsilon^\infty r^{1+\zeta} \frac{F_R(b_R(\lambda T \bar{F}_D(T))dr)}{\bar{F}_R(b_R(\lambda T \bar{F}_D(T)))} \to \alpha_R \int_\epsilon^\infty r^{\zeta-\alpha_R}dr = \frac{\alpha_R}{\alpha_R - 1 - \zeta}\epsilon^{1+\zeta-\alpha_R}, \quad T \to \infty.$$

Similarly, integration by parts and (2.27) with $T \geq T_0$ such that $\epsilon < T_0/T$ yields

$$\begin{aligned}\int_0^\epsilon u^{1+\zeta} \frac{F_D(Tdu)}{\bar{F}_D(T)} &= (1+\zeta)\int_0^\epsilon u^\zeta \frac{\bar{F}_D(Tu)}{\bar{F}_D(T)}du \\ &\leq (1+\zeta)\mu_D T_0^{\alpha_D - 1 + \eta} \int_0^\epsilon u^{\zeta - \alpha_D - \eta}du \\ &\leq \frac{(1+\zeta)\mu_D T_0^{\alpha_D-1+\eta}}{1+\zeta-\alpha_D-\eta}\epsilon^{1+\zeta-\alpha_D-\eta},\end{aligned}$$

where we used the bound (2.31). Thus

$$\limsup_{T\to\infty} \left|I_{(u<\epsilon, r>\epsilon)}\right| \leq constant \cdot \epsilon^{2(1+\zeta)-\alpha_R-\alpha_D-\zeta}, \quad (2.37)$$

where the exponent of ϵ is positive if we additionally let $(\alpha_R + \alpha_D + \eta)/2 < 1 + \zeta$.

For $I_{(r<\epsilon)}$, use again Lemmas 2.5.4 and 2.5.7 to get a $d_1 > 0$ such that

$$|I_{(r<\epsilon)}| \leq d_1 \sum_{j=1}^m \sum_{k=1}^m |z_j z_k| t_j \int_0^\epsilon r^2 \frac{F_R(b_R(\lambda T \bar{F}_D(T))dr)}{\bar{F}_R(b_R(\lambda T \bar{F}_D(T)))} \int_0^\infty u(t_k \wedge u) \frac{F_D(Tdu)}{\bar{F}_D(T)}.$$

Analogously to the bound for $I_{(u<\epsilon,r>\epsilon)}$, it can be readily shown that there exists $T_0 > 0$ such that for $T \geq T_0$

$$\int_0^\epsilon r^2 \frac{F_R(b_R(\lambda T \bar{F}_D(T))dr)}{\bar{F}_R(b_R(\lambda T \bar{F}_D(T)))} \leq \frac{3\mu_R T_0^{\alpha_R}}{2-\alpha_R} \epsilon^{2-\alpha_R},$$

and

$$\begin{aligned}\int_0^\infty u(t_k \wedge u) \frac{F_D(Tdu)}{\bar{F}_D(T)} &\leq \int_0^1 u^2 \frac{F_D(Tdu)}{\bar{F}_D(T)} + t_k \int_1^\infty u \frac{F_D(Tdu)}{\bar{F}_D(T)} \\ &\leq \frac{3\mu_D T_0^{\alpha_D}}{2-\alpha_D} + t_k \frac{\alpha_D}{\alpha_D - 1},\end{aligned}$$

whence

$$\limsup_{T\to\infty} \left| I_{(r<\epsilon)} \right| \leq constant \cdot \epsilon^{2-\alpha_R}. \tag{2.38}$$

Also, by fast growth and the regular variation of $\bar{F}_R$

$$\lambda T \bar{F}_D(T) \bar{F}_R(b_R(\lambda T \bar{F}_D(T))) \to 1, \qquad T \to \infty. \tag{2.39}$$

Finally, we can put together (2.35)-(2.39) to write:

$$\begin{aligned}\lim_{T\to\infty} \ln E \exp\left\{ i \sum_{j=1}^m z_j A_{cs}(t_j) \right\} &= \lim_{\epsilon\to 0} \lim_{T\to\infty} \ln E \exp\left\{ i \sum_{j=1}^m z_j A_{cs}(t_j) \right\} \\ &= \int_{-\infty}^\infty \int_0^\infty \int_0^\infty g(s,u,r) ds \cdot \alpha_D u^{-(\alpha_D+1)} du \cdot \alpha_R r^{-(\alpha_R+1)} dr.\end{aligned}$$

□

2.4 Remaining choices of α_D and α_R

We first study what happens if $\alpha_R < \alpha_D$, namely, $\alpha_R < \alpha_D^{(j)}$ for $j = 1, \ldots, p$. Consider the centered and scaled cumulative input of any of the streams, say

the first one, for simplicity. Set the normalizing term to $a^{(1)}(T) = b_R(\lambda T)$ (and the same for all other streams). Write

$$\begin{aligned} A_{cs}^{(1)}(t) &= \frac{1}{b_R(\lambda T)} \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} rL_{Tt}(s,u)\mathring{N}^{(1)}(ds,du,dr) \\ &= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} rL_{Tt}(Ts,u)\mathring{N}^{(1)}(Tds,du,b_R(\lambda T)dr). \end{aligned}$$

First, note from the definition of $L_t(s,u)$ in (2.8) that

$$\lim_{T\to\infty} L_{Tt}(Ts,u) = u1_{[0,t]}(s).$$

Thus, analogous to the proof of Proposition 2.5.6, it can be shown that the log fidi chf of $A_{cs}^{(1)}$ is

$$\begin{aligned} &\ln E\exp\left\{ i\sum_{j=1}^m z_j A_{cs}^{(1)}(t_j)\right\} \\ &= \frac{\lambda^{(1)}}{\lambda}\lambda T\bar{F}_R(b_R(\lambda T))\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} i\left(\exp\left\{ ir\sum_{j=1}^m z_j L_{Tt_j}(Ts,u)\right\} - 1\right) \\ &\times \sum_{k=1}^m z_k L_{Tt_j}(Ts,u)dsF_D^{(1)}(du)\frac{\bar{F}_R(b_R(\lambda T)r)}{\bar{F}_R(b_R(\lambda T))}dr. \end{aligned} \tag{2.40}$$

Now observe that, provided we can take the limit inside the integral

$$\begin{aligned} &\lim_{T\to\infty}\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} i\left(\exp\left\{ ir\sum_{j=1}^m z_j L_{Tt_j}(Ts,u)\right\} - 1\right)\times \\ &\qquad\qquad \sum_{k=1}^m z_k L_{Tt_j}(Ts,u)dsF_D^{(1)}(du)\frac{\bar{F}_R(b_R(\lambda T)r)}{\bar{F}_R(b_R(\lambda T))}dr \\ &= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty}\left(\exp\left\{ ir\sum_{j=1}^m z_j u1_{[0,t_j]}(s)\right\} - 1\right)\times \\ &\qquad\qquad \sum_{k=1}^m z_k u1_{[0,t_j]}(s)dsF_D^{(1)}(du)r^{-\alpha_R}dr. \end{aligned} \tag{2.41}$$

Since $|\lambda^{(1)}/\lambda| \leq 1$, then

$$\epsilon^{(1)}(T) := \ln E\exp\left\{ i\sum_{j=1}^m z_j A_{cs}^{(1)}(t_j)\right\} -$$

$$\frac{\lambda^{(1)}}{\lambda}\left\{\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} i\left(\exp\left\{ir\sum_{j=1}^{m} z_j u 1_{[0,t_j]}(s)\right\}-1\right)\times \right.$$
$$\left.\sum_{k=1}^{m} z_k u 1_{[0,t_j]}(s) ds F_D^{(1)}(du) r^{-\alpha_R} dr\right\} \to 0, \qquad T\to\infty, \tag{2.42}$$

which yields the following result.

Theorem 2.4.1. *Let $\Psi := \Psi_T$ be the fidi chf of*

$$\sum_{j=1}^{p}\frac{\lambda^{(j)}}{\lambda}E[(D_1^{(j)})^{\alpha_R}]^{1/\alpha_R}\Lambda_{\alpha_R}(t)$$
$$\stackrel{d}{=}\left(\left(-\cos\frac{\pi\alpha_R}{2}\right)\frac{2\Gamma(2-\alpha_R)}{\alpha_R(\alpha_R-1)}\right)^{-1/\alpha_R}\int_{-\infty}^{\infty}\int_{0}^{\infty} 1_{[0,t]}(s)uM_{\alpha_R}(ds,du), \tag{2.43}$$

where for each T, $\Lambda_{\alpha_R}(\cdot)$ is an α_R-stable Lévy motion totally skewed to the right with index α_R and $M_{\alpha_R}(ds,du)$ is α_R-stable with control measure $m(ds,du)=dsF_D(du)$. Then,

$$\lim_{T\to\infty}\left\{\ln E\exp\left\{i\sum_{j=1}^{m} z_j A_{cs}^{(1)}(t_j)\right\}-\ln\Psi_{t_1,\ldots,t_m}(z_1,\ldots,z_m)\right\}=0. \tag{2.44}$$

In addition, if for $j=1,\ldots,p$, the limits

$$w^{(j)} := \lim_{T\to\infty}\lambda^{(j)}/\lambda \tag{2.45}$$

exist, then the fidi chf of $A_{cs}(\cdot)$ converges to the fidi chf of the process defined by (2.43)*, with $w^{(j)}$ and $\sum_{j=1}^{p} w^{(j)}F_D^{(j)}(\cdot)$ replacing $\lambda^{(j)}/\lambda$ and $F_D(\cdot)$.*

Proof. By the independence of $N^{(j)}, j=1,\ldots,p$,

$$\ln E\exp\left\{i\sum_{j=1}^{m} z_j A_{cs}(t_j)\right\}$$
$$-\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} i\left(\exp\left\{ir\sum_{j=1}^{m} z_j u 1_{[0,t_j]}(s)\right\}-1\right)\sum_{k=1}^{m} z_k u 1_{[0,t_j]}(s)dsF_D(du)r^{-\alpha_R}dr$$
$$=\sum_{j=1}^{p}\epsilon^{(j)}(T)\to 0 \qquad T\to\infty. \tag{2.46}$$

Analogously to the proof of (2.59), the second integral in (2.46) is equal to $\ln \Psi_{t_1,\dots,t_m}(z_1,\dots,z_m)$. Thus, it only remains to justify taking the limit (2.42).

Let $0 < \zeta < \zeta' < 1$ and $0 < \eta < 1$ such that $1 + \zeta < \alpha_R - \eta$ and $\alpha_R + \eta < 1 + \zeta' < \alpha_D$. Similarly to (2.26) and (2.28), there exists $T_0 := T_0(\eta) > 0$ such that for $T \geq T_0$ and $b_R(\lambda T) \geq T_0$,

$$\frac{\bar{F}_R(b_R(\lambda T)r)}{\bar{F}_R(b_R(\lambda T))} \leq \begin{cases} 2r^{-\alpha_R}\left\{r^{-\eta} \vee r^{\eta}\right\}, & r \geq T_0/b_R(\lambda T), \\ \mu_R b_R(\lambda T)^{\alpha_R - 1 + \eta} r^{-1}, & r \in \mathbb{R}. \end{cases}$$

Together with Lemmas 2.5.4 and 2.5.7, this implies that the integrand in the left side of (2.41) is bounded in $\{r \geq 1\}$

$$B_{(>)} := 2^{1-\zeta} u^{1+\zeta} r^{\zeta - \alpha_R + \eta} \sum_{j=1}^{m} \sum_{k=1}^{m} |z_j|^{\zeta} |z_k| 1_{[0,t_k]}(s) 1_{[1,\infty)}(u),$$

and bounded in $\{r < 1\}$ by

$$B_{(<)} := 2^{1-\zeta'} \mu_R T_0^{\alpha_R - 1 + \eta} u^{1+\zeta'} r^{\zeta' - \alpha_R - \eta} \sum_{j=1}^{m} \sum_{k=1}^{m} |z_j|^{\zeta'} |z_k| 1_{[0,t_k]}(s) 1_{(0,1)}(r),$$

whenever $b_R(\lambda T) > T_0$. Here we used

$$r^{\zeta'} \leq (T_0/(b_R(\lambda T))^{\alpha_R - 1 + \eta} r^{1 + \zeta' - \alpha_D - \eta}.$$

By our choice of ζ, ζ' and η, both bounds are integrable and we can use dominated convergence to prove the result. □

In principle, it also is possible to have $\alpha_D = \alpha_R$. However, we cannot say much except in the special case $\alpha_D = \alpha_R = 2$, in which case the limit process is a Brownian motion provided (2.45) holds. We refer the reader to Kaj and Taqqu (2008, Theorem 4) for the formal statement of this case.

2.5 Technical proofs

This section contains a collection of technical results needed for our proofs. The first lemma establishes bounds for $b_D(\cdot) = (1/\bar{F}_D)^{\leftarrow}(\cdot)$ which yield $b_D(\lambda T) \to \infty$. This is not immediate since the function b_D depends on T.

Lemma 2.5.1. *The quantile functions given* (2.3) *and* (2.4) *satisfy the following inequality.*

$$\bigvee_{j=1}^{p} b_D^{(j)}(p\lambda^{(j)}T) \geq b_D(\lambda T) \geq \bigvee_{j=1}^{p} b_D^{(j)}(\lambda^{(j)}T), \quad T > 0. \tag{2.47}$$

Hence

$$b_D(\lambda T) \to \infty, \quad T \to \infty.$$

Proof. Since $\bar{F}_D^{(j)}$ is decreasing for all j, then

$$\begin{aligned}
\bar{F}_D\left(\bigvee_{j=1}^{p} b_D^{(j)}(p\lambda^{(j)}T)\right) &\leq \sum_{j=1}^{p} (\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(b_D^{(j)}(p\lambda^{(j)}T)) \\
&\leq \sum_{j=1}^{p} (\lambda^{(j)}/\lambda)(p\lambda^{(j)}T)^{-1} \\
&= (\lambda T)^{-1}.
\end{aligned}$$

Thus, the left side of (2.47) follows.

On the other hand, since F_D is right continuous, we have for each $j = 1, \ldots, p$:

$$(\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(b_D(\lambda T)) \leq \sum_{k=1}^{p} (\lambda^{(k)}/\lambda)\bar{F}_D^{(k)}(b_D(\lambda T)) \leq (\lambda T)^{-1},$$

whence

$$\bar{F}_D^{(j)}(b_D(\lambda T)) \leq (\lambda^{(j)}T)^{-1}.$$

Therefore, the right side of (2.47) follows. □

The distribution $F_D = \sum_{j=1}^{p}(\lambda^{(j)}/\lambda)F_D^{(j)}$ of session durations of superimposed streams is a function of T since $\lambda^{(j)}$ and λ depend on T. Nevertheless, $\bar{F}_D$ behaves as a regularly varying function.

Lemma 2.5.2. *Under the assumption* (2.23)

$$\lim_{T\to\infty}\frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} = \lim_{T\to\infty}\lambda T\bar{F}_D(b_D(\lambda T)u) = u^{-\alpha_D}, \quad u>0, \tag{2.48}$$

and therefore, in $M_+(0,\infty]$,

$$\frac{\bar{F}_D(Tdu)}{\bar{F}_D(T)} \xrightarrow{v} \alpha_D u^{-(\alpha_D+1)}du, \qquad T\to\infty. \tag{2.49}$$

Proof. Note that $\bar{F}_D \in RV_{-\alpha_D}$ for each fixed T. However, because F_D varies with T, the limit is not straightforward.

Fix an arbitrary $u>0$. We start by writing

$$\begin{aligned}\frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} &= \frac{\sum_{j:\alpha_D^{(j)}=\alpha_D}(\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(Tu)}{\bar{F}_D(T)} + \sum_{j:\alpha_D^{(j)}>\alpha_D}(\lambda^{(j)}/\lambda)\frac{\bar{F}_D^{(j)}(Tu)}{\bar{F}_D(T)}\\ &=: B + \sum_{j:\alpha_D^{(j)}>\alpha_D}(\lambda^{(j)}/\lambda)C_j,\end{aligned}$$

and additionally write

$$\begin{aligned}B^{-1} &= \frac{\sum_{j:\alpha_D^{(j)}=\alpha_D}(\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(T)}{\sum_{j:\alpha_D^{(j)}=\alpha_D}(\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(Tu)} + \sum_{j:\alpha_D^{(j)}>\alpha_D}(\lambda^{(j)}/\lambda)\frac{\bar{F}_D^{(j)}(T)}{\sum_{k:\alpha_k=\alpha_D}(\lambda_k/\lambda)\bar{F}_D^{(k)}(Tu)}\\ &=: B_1 + \sum_{j:\alpha_D^{(j)}>\alpha_D}(\lambda^{(j)}/\lambda)B_{2,j}.\end{aligned}$$

Thus, the first limit in (2.48) will follow by proving $B_1\to u^{\alpha_D}$, $B_{2,j}\to 0$ and $C_j\to 0$ as $T\to\infty$, for all j such that $\alpha_D^{(j)}>\alpha_D$.

First, by Potter's bounds applied to the regular variation of each $\bar{F}_D^{(j)}$, we have as $T \to \infty$,

$$B_1 \sim \frac{\sum_{j:\alpha_D^{(j)}=\alpha_D} (\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(T)}{\sum_{j:\alpha_D^{(j)}=\alpha_D} (\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(T)u^{-\alpha_D}} = u^{\alpha_D}.$$

Now, consider $B_{2,j}$ for $\alpha_D^{(j)} > \alpha_D$. Choose an arbitrarily large $z > \min\{u^{-\alpha_D^{(j)}}, u^{\alpha_D^{(j)}}\}$. By regular variation, for T sufficiently large:

$$\frac{\bar{F}_D^{(k)}(Tu)}{\bar{F}_D^{(j)}(Tu)} > z, \qquad \frac{\bar{F}_D^{(k)}(T)}{\bar{F}_D^{(j)}(T)} > z,$$

for all k such that $\alpha_D^{(k)} = \alpha_D$. In addition

$$\frac{\bar{F}_D^{(j)}(Tu)}{\bar{F}_D^{(j)}(T)} > u^{-a_D^{(j)}} - z^{-1}, \qquad \frac{\bar{F}_D^{(j)}(T)}{\bar{F}_D^{(j)}(Tu)} > u^{a_D^{(j)}} - z^{-1}.$$

Furthermore, the assumption (2.23) means there exists $d > 0$ such that for all T sufficiently large, there is some $k' := k'(T)$ such that $\alpha_D^{(k')} = \alpha_D$ and $\lambda^{(k')}/\lambda > d$. Hence for T sufficiently large:

$$\begin{aligned} B_{2,j}^{-1} &= \sum_{k:\alpha_D^{(k)}=\alpha_D} (\lambda^{(k)}/\lambda)\frac{\bar{F}_D^{(k)}(Tu)}{\bar{F}_D^{(j)}(T)} \\ &= \sum_{k:\alpha_D^{(k)}=\alpha_D} (\lambda^{(k)}/\lambda)\frac{\bar{F}_D^{(k)}(Tu)}{\bar{F}_D^{(j)}(Tu)}\frac{\bar{F}_D^{(j)}(Tu)}{\bar{F}_D^{(j)}(T)} \\ &> dz(u^{-\alpha_D} - z^{-1}). \end{aligned}$$

This shows that $B_{2,j}^{-1}$ can be made arbitrarily large for T sufficiently large, whence $B_{2,j} \to 0$ as $T \to \infty$.

Similarly, consider C_j, and

$$\begin{aligned} C_j^{-1} &\geq \sum_{k:\alpha_D^{(k)}=\alpha_D} (\lambda^{(k)}/\lambda) \frac{\bar{F}_D^{(k)}(T)}{\bar{F}_D^{(j)}(Tu)} \\ &= \sum_{k:\alpha_D^{(k)}=\alpha_D} (\lambda^{(k)}/\lambda) \frac{\bar{F}_D^{(k)}(T)}{\bar{F}_D^{(j)}(T)} \frac{\bar{F}_D^{(j)}(T)}{\bar{F}_D^{(j)}(Tu)} \\ &> dz(u^{\alpha_D} - z^{-1}). \end{aligned}$$

This shows that C_j^{-1} can be made arbitrarily large for T sufficiently large, which completes the first part of the Lemma.

For the second limit in (2.48), recall that $z < b_D(\lambda T)$ iff $1/\bar{F}_D(z) < \lambda T$ for each T. For $\epsilon > 0$, setting $z = b_D(\lambda T)(1-\epsilon)$ and $z = b_D(\lambda T)(1+\epsilon)$ yields

$$\frac{\bar{F}_D(b_D(\lambda T)(1+\epsilon))}{\bar{F}_D(b_D(\lambda T))} \leq \frac{1}{\lambda T \bar{F}_D(b_D(\lambda T))} \leq \frac{\bar{F}_D(b_D(\lambda T)(1-\epsilon))}{\bar{F}_D(b_D(\lambda T))}.$$

Letting $T \to \infty$ and using Lemma 2.5.1 and the first limit gives

$$(1+\epsilon)^{-\alpha_D} \leq \frac{1}{\lambda T \bar{F}_D(b_D(\lambda T))} \leq (1-\epsilon)^{-\alpha_D}.$$

Because ϵ is arbitrary, then

$$\lim_{T\to\infty} \lambda T \bar{F}_D(b_D(\lambda T)) = 1.$$

Therefore

$$\lim_{T\to\infty} \lambda T \bar{F}_D(b_D(\lambda T)u) = \lim_{T\to\infty} \lambda T \bar{F}_D(b_D(\lambda T)) \lim_{T\to\infty} \frac{\bar{F}_D(b_D(\lambda T)u)}{\bar{F}_D(b_D(\lambda T))} = u^{-\alpha_D}.$$

The final statement about vague convergence follows the proof of Resnick (2007, Theorem 3.6). □

Even though F_D depends on T, a version of Potter's bounds holds.

Lemma 2.5.3. *Let $\delta > 0$. Under the assumption* (2.23)*, there exists $T_0 = T_0(\delta) > 0$ such that for all $T \geq T_0$, $Tu \geq T_0$:*

$$\frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} \leq (1+\delta)u^{-\alpha_D} \max\{u^{-\delta}, u^{\delta}\}.$$

Proof. Observe

$$\frac{\bar{F}_D(Tu)}{\bar{F}_D(T)} = \frac{\sum_{j=1}^{p}(\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(T)\frac{\bar{F}_D^{(j)}(Tu)}{\bar{F}_D^{(j)}(T)}}{\sum_{j=1}^{p}(\lambda^{(j)}/\lambda)\bar{F}_D^{(j)}(T)} \leq \bigvee_{j=1}^{p} \frac{\bar{F}_D^{(j)}(Tu)}{\bar{F}_D^{(j)}(T)}.$$

By Potter bounds (See e.g. Bingham et al., 1987, Theorem 1.5.6), for all $j = 1, \ldots, p$ there exists $T_j = T_j(\delta)$ such that

$$\frac{\bar{F}_D^{(j)}(Tu)}{\bar{F}_D^{(j)}(T)} \leq (1+\delta)u^{-\alpha_D^{(j)}} \max\{u^{-\delta}, u^{\delta}\} \leq (1+\delta)u^{-\alpha_D} \max\{u^{-\delta}, u^{\delta}\},$$

for $T \geq T_j$, $Tu \geq T_j$. Therefore, the result holds for $T_0 = \bigvee_{j=1}^{p} T_j$. □

We now study $L_t(s,u)$, as defined in (2.8).

Lemma 2.5.4. *The length of the subinterval of $[0,t]$ during which the session (s,u,r) transmits data, namely $L_t(s,u)$ in* (2.8)*, satisfies the following properties:*

(i) ***Scaling property:*** *For $C > 0$,*

$$CL_t(s,u) = L_{Ct}(Cs, Cu).$$

(ii) ***Bounds:***

$$L_t(s,u) \leq t \wedge u. \tag{2.50}$$

(iii) ***Integrals:*** *For $1 < \gamma < 2$ and nonnegative t_1, t_2,*

$$\int_{-\infty}^{\infty} L_{t_1}(s,u)ds = ut_1,$$

and

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} L_{t_1}(s-u,u)1_{[0,t_2]}(s)u^{-\gamma}duds$$
$$= \frac{1}{(\gamma-1)(2-\gamma)(3-\gamma)}\left\{(t_2^{3-\gamma}-(t_2-t_1)^{3-\gamma})1_{t_1<t_2}+t_2^{3-\gamma}1_{t_1\geq t_2}\right\}.$$

Proof. The scaling property and the bounds follow directly from (2.8).

Now, the first part of Property (iii) is readily checked by using the first integral in (2.8) after reversing the order of integration. Finally, the second part of Property (iii) can be derived by writing

$$L_{t_1}(s-u,u) = \begin{cases} 0, & s<0 \text{ or } s>u+t_1, \\ s, & 0\leq s\leq u\wedge t_1, \\ t_1, & t_1\leq s\leq u, \\ u, & u\leq s\leq t_1, \\ t_1-s+u, & u\vee t_1\leq s\leq u+t_1, \end{cases}$$

and integrating accordingly. Observe that the four regions in which $L_{t_1}(s-u,u)$ is nonzero correspond to those of the basic decomposition in Mikosch et al. (2002, Equation 4.1).

□

The next lemma helps obtain approximations to the cumulative input of the aggregated streams.

Lemma 2.5.5. *For any $a, T > 0$, we have*

$$\begin{aligned}\frac{1}{a}(A(Tt) - \lambda\mu_D\mu_R Tt) &= \frac{1}{a}\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} rL_{Tt}(Ts,u)\mathring{N}(Tds,du,dr)\\ &= \frac{T}{a}\int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} rL_t(s,u)\mathring{N}(Tds,Tdu,dr)\\ &= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} rL_t(s,u)\mathring{N}(Tds,Tdu,(a/T)dr).\end{aligned}$$

Proof. All the relations here follow from several ways to change variables in (2.13) and using the scaling property of $L_t(s,u)$. See Lemma 2.5.4. □

Our limit theorems are proved by verifying convergence of finite dimensional distributions for various processes. The following is required.

Proposition 2.5.6. *For arbitrary $m \geq 1$, $0 \leq t_1, \ldots, t_m$, and real $z_1, \ldots, z_m$, define*

$$g(s,u,r) = \exp\left\{i\sum_{j=1}^m z_j rL_{t_j}(s,u)\right\} - 1 - i\sum_{j=1}^m z_j rL_{t_j}(s,u), \tag{2.51}$$

and

$$h(s,u,r) = i\left(\exp\left\{i\sum_{j=1}^m z_j ur1_{(0,t_j)}(s)\right\} - 1\right)\sum_{k=1}^m z_k 1_{[0,t_k]}(s) ru^{-\alpha_D}. \tag{2.52}$$

(a) For any $a, T > 0$, the characteristic function of the finite-dimensional distributions (fidi chf) *of the process $\{(1/a)(A(Tt) - \lambda\mu_D\mu_R Tt); t \geq 0\}$ is given by*

$$\begin{aligned}&\ln E\exp\left\{i\sum_{j=1}^m z_j\left[\frac{1}{a}(A(Tt_j) - \lambda\mu_D\mu_R Tt_j)\right]\right\}\\ &\quad= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} g(s,u,r)EN(Tds,Tdu,(a/T)dr) \qquad (2.53)\\ &\quad= \int_{-\infty}^{\infty}\int_0^{\infty}\int_0^{\infty} g_u(s-u,u,r)\lambda T\bar{F}_D(Tu)dsduF_R((a/T)dr). \qquad (2.54)\end{aligned}$$

where g_u is the partial derivative of g with respect to u.

(b) The fidi chf of the limit processes in Corollary 2.2.2 are given as follows.

(i) The fidi chf of the limit process under Scenario $\mathcal{F}$ and $E[(R_1^{(1)})^2] < \infty$ is given by

$$\begin{aligned}\ln E \exp\left\{ i \sum_{j=1}^{m} z_j E[(R_1^{(1)})^2]^{1/2} \sigma_{B_H(1)} B_H(t_j) \right\} \\ = -\frac{1}{2} E[(R_1^{(1)})^2] \sum_{j=1}^{m} \sum_{k=1}^{m} z_i z_j \sigma^2_{B_H(1)} \frac{1}{2} \left(t_i^{2H} + t_j^{2H} - |t_i - t_j|^2 \right),\end{aligned} \tag{2.55}$$

where B_H is fractional Brownian motion with

$$\sigma^2_{B_H(1)} = \frac{2}{(\alpha_D - 1)(2 - \alpha_D)(3 - \alpha_D)}, \tag{2.56}$$

and $H = (3 - \alpha_D)/2$.

(ii) The fidi chf of the limit process under Scenario $\mathcal{F}$ and $\bar{F}_R \in RV_{-\alpha_R}$, $1 < \alpha_R < 2$, is given by

$$\begin{aligned}\ln E \exp\left\{ i \sum_{j=1}^{m} z_j Z_{\alpha_D, \alpha_R}(t_j) \right\} \\ = \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{\infty} g(s,u,r) E N^{\infty}_{\alpha_D, \alpha_R}(ds, du, dr).\end{aligned} \tag{2.57}$$

(iii) The fidi chf of the limit process under Scenario $\mathcal{M}$ is given by

$$\begin{aligned}\ln E \exp\left\{ i \sum_{j=1}^{m} z_j c Y_{\alpha_D}(t_j/c) \right\} \\ = c^{\alpha_D - 1} \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{\infty} g(s,u,r) E N^{\infty}_{\alpha_D, F_R}(ds, du, dr).\end{aligned} \tag{2.58}$$

(iv) Finally, the fidi chf of the limit process under Scenario $\mathcal{S}$ is given by

$$\begin{aligned}\ln E \exp\left\{ i \sum_{j=1}^{m} z_j E[(R_1^{(1)})^{\alpha_D}]^{1/\alpha_D} \Lambda_{a_D}(t_j) \right\} \\ = \int_{-\infty}^{\infty} \int_{0}^{\infty} \int_{0}^{\infty} h(s,u,r) ds du F_R(dr).\end{aligned} \tag{2.59}$$

Proof. Given (2.53), (2.54) is readily derived using integration by parts and the change of variables $s \mapsto s + u$. Moreover, (2.55) follows from the fact that B_H is fractional Brownian motion. The remaining parts are a consequence of the following property of Poisson random measures (See e.g. Rosiński and Rajput, 1989):

$$\ln E \exp\left\{ i \int f(x) \mathring{\xi}(dx) \right\} = \int \left(e^{if(x)} - 1 - if(x) \right) E\xi(dx),$$

if

$$\int \left(f^2(x) \wedge |f(x)| \right) E\xi(dx) < \infty.$$

For now, let us focus on (2.53), (2.57) and (2.58). By Lemma 2.5.5, the exponent in the left side of (2.53) and (2.57) is of the form

$$i \int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} \sum_{j=1}^{m} z_j r L_{t_j}(s,u) \mathring{\xi}(ds, du, dr), \tag{2.60}$$

for a $\mathbb{PRM}$ ξ, while the exponent in the left side of (2.58) is $c^{\alpha_D - 1}$ times (2.60), using Lemma 2.5.4 and the change of variables $s \mapsto s/c, u \mapsto u/c$. Thus, it suffices to check that

$$\int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} \left(\sum_{j=1}^{m} z_j r L_{t_j}(s,u) \right)^2 \bigwedge \left| \sum_{k=1}^{m} z_k r L_{t_k}(s,u) \right| E\xi(ds, du, dr) < \infty. \tag{2.61}$$

Bounds and integral results for $L_t(s,u)$ in Lemma 2.5.4 (ii) and (iii) needed.

First observe that

$$\int_{-\infty}^{\infty} \int_0^{\infty} \int_0^{\infty} \left| \sum_{j=1}^{m} z_j r L_{t_j}(s,u) \right| EN(Tds, Tdu, (a/T)dr) \leq \frac{T}{a} \sum_{j=1}^{m} |z_j| \lambda \mu_D \mu_R t_j,$$

which proves (2.53).

In order to prove (2.57), split the corresponding integral (2.61) into two parts $I_{(<)}$ and $I_{(>)}$, according to the two domains of integration $D_{(<)} = \{ur < 1\}$ and

$D_{(>)} = \{ur > 1\}$. This yields

$$\begin{aligned} I_{(<)} &\le \sum_{j=1}^{m}\sum_{k=1}^{m} |z_j z_k| \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{1/u} r^2 L_{t_j}(s,u) L_{t_k}(s,u) E N^{\infty}_{\alpha_D,\alpha_R}(ds,du,dr) \\ &\le \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{|z_j z_k| t_j \alpha_D \alpha_R}{2-\alpha_R} \left(\frac{1}{\alpha_R - \alpha_D} + \frac{t_k}{1-\alpha_R+\alpha_D} \right), \end{aligned}$$

and

$$\begin{aligned} I_{(>)} &\le \sum_{j=1}^{m} |z_j| \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{1\vee u^{-1}}^{\infty} r L_{t_j}(s,u) E N^{\infty}_{\alpha_D,\alpha_R}(ds,du,dr) \\ &\quad + \sum_{j=1}^{m}\sum_{k=1}^{m} |z_j z_k| \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{u^{-1}}^{1\vee u^{-1}} r^2 L_{t_j}(s,u) L_{t_k}(s,u) E N^{\infty}_{\alpha_D,\alpha_R}(ds,du,dr) \\ &\le \sum_{j=1}^{m} \frac{|z_j| t_j \alpha_D \alpha_R}{\alpha_R - 1} \left(\frac{1}{\alpha_R - \alpha_D} + \frac{1}{\alpha_D} \right) + \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{|z_j z_k t_j t_k|}{(\alpha_D - 1)(2-\alpha_R)}, \end{aligned}$$

whence (2.57) holds.

Similarily, split the integral (2.61) corresponding to the process (2.58) into two parts $J_{(<)}$ and $J_{(>)}$, according to the two domains of integration $D_{(<)}$ and $D_{(>)}$, which yields

$$\begin{aligned} J_{(<)} &\le \sum_{j=1}^{m}\sum_{k=1}^{m} |z_j z_k| \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{r^{-1}} r^2 L_{t_j}(s,u) L_{t_k}(s,u) E N^{\infty}_{\alpha_D,F_R}(ds,du,dr) \\ &\le \sum_{j=1}^{m}\sum_{k=1}^{m} \frac{|z_j z_k| t_j \alpha_D}{2-\alpha_D} E[(R_1^{(1)})^{\alpha_D}], \end{aligned}$$

and

$$\begin{aligned} J_{(>)} &\le \sum_{j=1}^{m} |z_j| \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{r^{-1}}^{\infty} r L_{t_j}(s,u) E N^{\infty}_{\alpha_D,F_R}(ds,du,dr) \\ &\le \sum_{j=1}^{m} \frac{|z_j| t_j \alpha_D}{\alpha_D - 1} E[(R_1^{(1)})^{\alpha_D}]. \end{aligned}$$

This proves (2.58).

Finally, the exponent in the left side of (2.59) is

$$i \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} \sum_{j=1}^{m} z_j u r 1_{[0,t_j]}(s) \mathring{N}^{\infty}_{\alpha_D,F_R}(ds,du,dr).$$

Analogous to the proof of (2.58), it is readily shown that

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty}\left(\sum_{j=1}^{m} z_j u r 1_{[0,t_j]}(s)\right)^2 \bigwedge \left|\sum_{k=1}^{m} z_k u r 1_{[0,t_j]}(s)\right| EN^{\infty}_{\alpha_D,F_R}(ds,du,dr) < \infty,$$

whence the left side of (2.59) is equal to

$$\int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty}\left(\exp\left\{i\sum_{j=1}^{m} z_j u r 1_{[0,t_j]}(s)\right\} - 1 - \sum_{j=1}^{k} z_j u r 1_{[0,t_j]}(s)\right) \times EN^{\infty}_{\alpha_D,F_R}(ds,du,dr).$$

The result now follows after an integration by parts in the variable u. □

Finally, the following result is used to get upper bounds for some integrands throughout the proof of Theorem 2.3.1.

Lemma 2.5.7. *For $0 \le \zeta \le 1$ and $x \in \mathbb{R}$:*

$$|e^{ix} - 1| \le 2^{1-\zeta}|x|^{\zeta}, \tag{2.62}$$

$$|e^{ix} - 1 - ix| \le d_{\zeta}|x|^{\zeta+1}, \tag{2.63}$$

where $d_\zeta > 0$, and for real numbers $x_1, \ldots, x_m$:

$$\left(\sum_{j=1}^{m}|x_j|\right)^{\zeta} \le \sum_{j=1}^{m}|x_j|^{\zeta}. \tag{2.64}$$

Proof. Without loss of generality, fix $x \neq 0$. Define $f : [0,1] \to \mathbb{R}$, $f(\zeta) = (1-\zeta)\ln 2 + \zeta \ln|x|$. We can readily check that

$$\ln|e^{ix} - 1| \le f(\zeta), \quad \zeta = 0, 1,$$

by taking logarithms in both sides of $|e^{ix} - 1| \le 2 \wedge |x|$. Since $f(\zeta)$ is linear in ζ, $f(\zeta)$ is either nondecreasing or nonincreasing on $[0,1]$. Hence, (2.62) holds. Using a similar strategy, we can prove (2.63).

For (2.64), assume without loss of generality that $0 < |x_1| \leq |x_2|$, thus $0 < |x_1/x_2| \leq 1$. By Bernoulli's inequality (see e.g. Mitrinović and Vasić, 1970, p. 36):

$$(1 + |x_1/x_2|)^\zeta \leq 1 + \zeta|x_1/x_2| \leq 1 + |x_1/x_2|^\zeta.$$

Multiplying both sides by $|x_2|^\zeta$ proves (2.64) for $m = 2$ and the proof for general m follows by induction. □

CHAPTER 3

EXTREMAL DEPENDENCE ANALYSIS OF NETWORK SESSIONS

3.1 Overview

This chapter is motivated by the study of two network invariants:

- Heavy tails for quantities such as file sizes (Leland et al., 1994; Willinger et al., 1998; Arlitt and Williamson, 1996; Willinger and Paxson, 1998), transmission durations and transmission delays (Maulik et al., 2002; Resnick, 2003).
- Bursty network traffic (Sarvotham et al., 2005), with rare but influential periods of high transmission rate punctuating typical periods of modest activity.

When studying burstiness, bursts are observed in the sequence of bytes-per-time or packets-per-time, which means that a window resolution is selected and the number of bytes or packets is counted over consecutive windows. Sarvotham et al. (2005) attempt to explain the causes of burstiness at the user-level. If the primary objective is to explain sources of burstiness, the session peak rate is a variable of interest in addition to the usual session descriptors of size, duration, and average transfer rate as defined in Section 1.1. The peak rate is computed as the maximum transfer rate over consecutive time slots within a session.

In order to explain the causes of burstiness at the user-level, Sarvotham et al. (2005) studied the dependence structure of quantities such as session size, du-

ration and transfer rate. They concluded that it is useful to split the data into two groups according to the values of peak rate and consider the properties of each group. These two groups were called alpha sessions consisting of sessions whose peak rate is above a high quantile, and beta sessions, comprising the remaining traffic. Various criteria for segmenting into the two groups were considered but a common theme was that the alpha sessions corresponded to "power users" who transmit large files at large bandwidth, and the beta sessions were the remaining ones. This analysis yielded the following:

- A tiny alpha group relative to a huge beta group. In addition, it appeared that the alpha group was the major source of burstiness.
- A dependence structure that is quite different in the alpha and beta groups, with approximate independence between rate and size for the alpha group and approximate independence between rate and duration for the beta group. To see this, Sarvotham et al. (2005) measured dependence with correlations between the log-variables.

We wondered if the large beta group should be treated as one homogeneous collection of users, especially when one is happy to identify a small and distinct alpha group. Thus, we have investigated whether segmenting the beta group further produces meaningful information.

Section 3.2 contains more details on the network traffic traces that we study, and gives the precise definition of session, size, duration, rate and peak rate. Historically (Crovella and Bestavros, 1997; Leland et al., 1994; Willinger et al., 1995, 1997), data collection was over finely resolved time intervals, and thus a natural definition of peak rate is based on computing the maximum transfer

rate over consecutive time slots. We discuss in Section 3.2 that this definition may be flawed due to the choice of the time window resolution giving the peak rate undesired properties. Thus we propose our own definition of peak rate.

In Section 3.3, we study the marginal distributions of size, duration and rate, and in Sections 3.4 and 3.5 we explore the dependence structure between these three variables. Throughout these sections, we depart from the approach of Sarvotham et al. (2005) by not just looking at the alpha and beta groups; instead, we have split the data into q groups of approximately equal size according to the quantiles of peak rate. Thus, where we previously had a beta group, we now have $q-1$ groups, whose peak rates are in a fixed quantile range. We show that the alpha/beta split is masking further structure and that it is important to take into account the explicit level of the peak rate. In Sections 3.4 and 3.5, we also review and use methods that are more suitable than correlation in the context of heavy tailed-modeling for studying the dependence of two variables.

We also have considered in Section 3.6 whether session starting times can be described by a Poisson process. While several authors have shown that the process of packet arrivals to servers cannot be modeled under the framework of Poisson processes (Paxson and Floyd, 1995; Willinger et al., 1997; Willinger and Paxson, 1998; Hohn et al., 2003), some argue that the network traffic is driven by independent human activity and thus justify the search for this underlying Poisson structure at higher levels of aggregation (Park et al., 2006). We found that while the homogeneous Poisson process cannot describe overall network traffic, it is a good model for session initiation times within each of the q groups produced by our segmentation of the overall traffic. In Section 3.7 we give some remarks including a rough outline for simulation of data sets based on the afore-

mentioned Poisson framework, and give possible lines of future study.

3.2 Definitions

3.2.1 Size S, duration D, and rate R of e2e sessions

In this chapter, we follow Sarvotham et al. (2005); Willinger et al. (1997)'s approach to define an *end to end (e2e) session*, or briefly session, as a cluster of bytes with the same source and destination network addresses, such that the delay between any two successive packets in the cluster is less than a threshold t. A session plays the role of an arriving entity in an infinite-source Poisson model.

For each session, we have the following variables:

- S represents the size, that is, the number of bytes transmitted.
- D represents the duration, computed as the difference in seconds between the arrival times of the first and last packets in the session.
- R represents the average transfer rate, namely S/D.

Note that R is not defined for single-packet sessions, for which D by definition is zero. More generally, sessions with very small D may also be problematic to handle. For instance, it would be hard to believe that a session sending only two packets back-to-back has an R that equals the line bandwidth. In order to avoid this issue, for our analysis we ignore sessions with $D < 100ms$. See Zhang et al. (2002) for related comments.

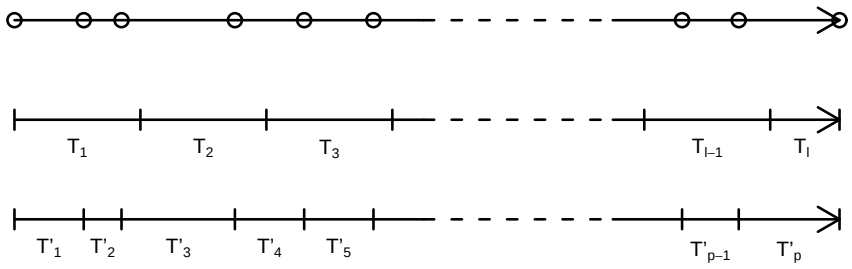

Figure 3.1: *Top arrow:* Representation of a typical session; here each packet is depicted as an oval. *Middle arrow:* Sarvotham et al. (2005)'s division approach. *Bottom arrow:* Our proposed division according to the packet arrival times.

3.2.2 Predictors of burstiness

In addition to S, D and R, Sarvotham et al. (2005) consider a fourth quantity which serves as an explanatory variable for burstiness, namely the session's maximum input in consecutive time windows. A closely related variable arises by considering the session's peak rate in consecutive intervals. In what follows, we review the properties of these two variables and show that they are not ideal for describing burstiness. Therefore, we propose and use a different definition of peak rate.

The δ-maximum input.

Fix a small $\delta > 0$ and divide each session in l subintervals of length δ, where $l = \lceil D/\delta \rceil$ (see Figure 3.1, Top and Middle). For $i = 1, \ldots, l$, define the following auxiliary variables:

- B_i represents the number of bytes transmitted over the ith subinterval of the session.
- T_i represents the duration of the ith subinterval. For $i = 1, \ldots, l-1$, we have $T_i = \delta$. However, notice that $T_l = D - (l-1)\delta$.

The *δ-maximum input* of the session is defined as $I_\delta = \bigvee_{i=1}^{n} B_i$. This I_δ is the original variable used by Sarvotham et al. (2005).

The δ-peak rate.

If the goal is to explain burstiness, a natural alternative to maximum input is to consider rates in consecutive time subintervals, rather than inputs. This yields a closely related predictor: the δ-peak rate. The definition of the *δ-peak rate* for a session, denoted as R_δ, relies on the Sarvotham et al. (2005)'s division of the session (see Figure 3.1). We define $R_\delta = \bigvee_{i=1}^{n} B_i/T_i$.

Observe the following properties for a session:

(i) $\sum_{i=1}^{n} B_i = S$;

(ii) $\sum_{i=1}^{n} T_i = D$;

(iii) $R_\delta \geq R$. To see this, note

$$R = \frac{S}{D} = \frac{\sum_{i=1}^{n} T_i \cdot \frac{B_i}{T_i}}{\sum_{i=1}^{n} T_i} \leq \bigvee_{i=1}^{n} B_i/T_i = R_\delta.$$

A quick analysis shows that the last property does not necessarily hold if we do not carefully define the duration of the last subinterval T_n as above, but instead set $T_n = \delta$. For a numerical example, let $\delta = 1$ and consider a session

with $n = 2, B_1 = B_2 = 1, D = 1.1$. Using the wrong definition $T_n = \delta$ yields $T_1 = T_2 = 1$, hence the average transfer rate $R = (B_1 + B_2)/D = 2/1.1$ but the peak transfer rate $R_\delta = \max\{B_1/T_1, B_2/T_2\} = 1 < 2/(1.1)$.

While both I_δ and R_δ appear to be natural predictors of burstiness, they both possess undesirable properties. They both depend on the parameter δ which is not an intrinsic characteristic of the session. As $\delta \downarrow 0$, many consecutive subintervals thus have a single packet, as in Figure 3.1. Therefore, as $\delta \downarrow 0$,

- $I_\delta \rightarrow$ *maximum packet size*, which precludes I_δ from being a useful measure of burstiness.
- $R_\delta \rightarrow \infty$, implying that R_δ is greater than the line capacity for small δ. Depending on the relationship of packet arrivals to the size of δ, we can get unreasonably large R_δs and therefore, the interpretation of R_δ as peak transfer rate becomes problematic.

Owing to the drawbacks of the previous two definitions, we propose our own definition of peak rate.

Peak rate $R^\vee$.

Suppose a session has p packets (see Figure 3.1, Bottom). Consider the following variables.

- B_i' represents the number of bytes of the ith packet.
- T_i' represents the interarrival time of the ith and $(i+1)$th packets, $i = 1, \ldots, p-1$.

For $k = 2, \ldots, p$, we define the *peak rate of order* k, denoted by $R^{(k)}$, as

$$R^{(k)} = \bigvee_{j=1}^{p-k+1} \frac{\sum_{i=j}^{j+k-1} B'_i}{\sum_{i=j}^{j+k-2} T'_i}. \tag{3.1}$$

In the above definition, the quotient measures the actual transfer rate of a stream of bytes consisting of k consecutive packets. For a session consisting of p packets, there are $p - k + 1$ streams of k consecutive packets, hence $R^{(k)}$ is a measure of the actual peak transfer rate when only k consecutive packets are taken into account. We then define the *peak rate* as

$$R^{\vee} = \bigvee_{k=2}^{p} R^{(k)}. \tag{3.2}$$

Notice that $R^{\vee} \geq R^{(p)} = R$. Moreover, one readily checks that there is always a neighboring packet pair whose rate is no less than the rate of any k consecutive packets that include the pair, whence $R^{\vee} = R^{(2)}$.

As opposed to I_δ and R_δ, $R^{\vee}$ does not depend on an external parameter δ, and thus it is an intrinsic characteristic of a session. In addition, $R^{\vee}$ inherits the interpretation of $R^{(k)}$ and therefore may be interpreted as a measure of the maximum transfer rate over all possible streams of consecutive packets. In addition, $R^{\vee} = R^{(2)}$ gives a simpler representation useful for computing $R^{\vee}$ in practice which is also suitable for analytical studies.

3.2.3 The data set

We present our results for an anonymized network trace captured at the University of Auckland between December 7 and 8, 1999, which was publicly available as of June 2011 through the Réseaux IP Européens (French for European IP Networks) Network Coordination Centre's data repository at `http:`

//labs.ripe.net/datarepository/. Auckland's data set is a collection of GPS-synchronized traces, where all non-IP traffic has been discarded and only TCP, UDP and ICMP traffic is present in the trace. We have taken the part of the trace corresponding exclusively to incoming TCP traffic sent on December 8, 1999, between 3 and 4 p.m. We have found that our results hold for several other data sets. See Section 3.7 for more details about this and other data traces.

The raw data consists of 1,177,497 packet headers, from which we construct 44,136 sessions using a threshold between sessions of $t = 2s$ and considering only those sessions with $D > 100ms$ (as explained in the last paragraph of Section 3.2.1). We have found similar results for various choices of thresholds between sessions, including $t = 0.1, 0.5, 10, 60, 100s$, but here we only present our results for $t = 2s$.

In addition, for each session we have peak rate $R_i^{\vee}$ and starting time Γ_i. Thus, the data set has the form $\{((S_i, D_i, R_i), R_i^{\vee}, \Gamma_i); 1 \leq i \leq 44,136\}$, that is, a set of 5-tuples. We are interested in the dependence structure of triplet (S_i, D_i, R_i).

We split these sessions into 10 groups of approximately equal size according to the empirical deciles of $R^{\vee}$. Thus, all the sessions in the gth group, $g = 1, \ldots, 10$, have $R^{\vee}$ is in a fixed decile range, $(10(g-1)\%, 10g\%]$. Hence we term the group of sessions "the gth decile group", $g = 1, \ldots, 10$. Therefore, where Sarvotham et al. (2005) had alpha and beta groups, we now have a more refined segmentation.

In the remainder of the chapter, we show that this refined split reveals features that are hidden by an elementary alpha/beta split.

3.3 Marginal distributions of S, D and R.

We analyze marginal distributions of S, D and R in the 10 different decile groups to check if heavy tails are present. For all decile groups, S and D have heavy tails, but unsurprisingly, not R.

3.3.1 Domain of attraction diagnostics

Excesses over high thresholds

Using the POT method, we found no evidence against $F_S, F_D \in \mathcal{D}(G_\gamma)$ for all the 10 decile groups. Typical QQ plots are those corresponding to the $GPD_{\gamma,\beta}$ fit for the excess of S and D in the 10th decile group, shown in Figure 3.2, *upper left* and *upper right* panels, respectively. Both plots exhibit an almost perfect straight line. We also found that the QQ plots corresponding to the excesses of S and D in all the other decile groups exhibit straight line trends, thus showing no evidence against satisfaction of the extreme value condition.

Similarly, Figure 3.2 *lower left* panel exhibits the QQ plot of the $GPD_{\gamma,\beta}$ fit for the log-excesses of R in the 10th decile group, which shows no evidence against $F_R \in \mathcal{D}(G_\gamma)$. However, for all the other decile groups, we found evidence against $F_R \in \mathcal{D}(G_\gamma)$. For instance, Figure 3.2 *Lower right* panel shows a QQ plot of the $GPD_{\gamma,\beta}$ fit for the log-excesses of R in the 4th decile group, exhibiting a departure from the straight line. We also found no straight line trend in the rest of the QQ plots of the $GPD_{\gamma,\beta}$ fit for the log-excesses of R in the lower nine decile groups.

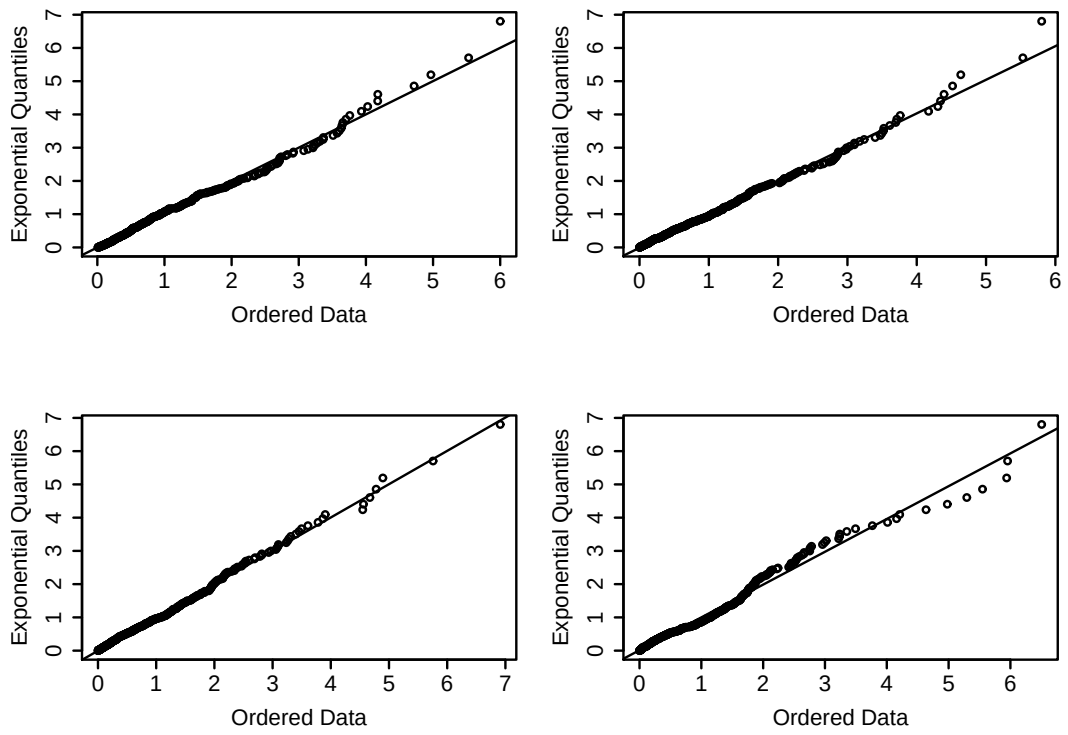

Figure 3.2: GPD QQ plots of excesses; a number $k = 450$ of upper order statistics is used for each fit. *Upper left:* Size in the 10th decile group. *Upper right:* Duration in the 10th decile group. *Lower left:* Rate in the 10th decile group. *Lower right:* Rate in the 4th decile group.

Formal tests of domain of attraction

Recently, two formal methods for testing $F \in \mathcal{D}(G_\gamma)$ have been derived by Dietrich et al. (2002) and Drees et al. (2006); see also de Haan and Ferreira (2006). Both tests are based on quantile function versions of the well known Crámer von-Mises and Anderson-Darling test statistics (see e.g. Lehmann and Romano, 2005), respectively, for checking the goodness of fit of a given distribution. In addition, both tests assume a second order condition which is difficult to check in practice. A follow-up study by Hüsler and Li (2006) examines the two tests'

error and power by simulations. A thorough discussion of these tests and the second order condition is provided by de Haan and Ferreira (2006). Here we review and apply the method.

Dietrich et al. (2002) state the following: Suppose $Y_1, \ldots, Y_n$ are iid with common distribution F and $Y_{1:n} \leq Y_{2:n} \leq \cdots \leq Y_{n:n}$ are the order statistics. If $F \in \mathcal{D}(G_\gamma)$ for some $\gamma \in \mathbb{R}$ and also F satisfies an additional second order tail condition (see Dietrich et al., 2002, equation (4)), then:

$$\begin{aligned} E_{k,n} :=& k \int_0^1 \left(\frac{\ln Y_{n-[kt]:n} - \ln Y_{n-k:n}}{\hat{\gamma}_+} - \frac{t^{-\hat{\gamma}_-} - 1}{\hat{\gamma}_-} \right) t^2 dt \\ \xrightarrow{d} E_\gamma :=& \int_0^1 \Big((1-\gamma_-)(t^{-\gamma_- - 1} W(t) - W(1)) - (1-\gamma_-)^2 \frac{t^{-\gamma_-} - 1}{\gamma_-} P_{\gamma_-} \\ &+ \frac{t^{-\gamma_-} - 1}{\gamma_-} R_{\gamma_-} + (1-\gamma_-) R_{\gamma_-} \int_t^1 s^{-\gamma_- - 1} \ln s ds \Big)^2 t^2 dt, \end{aligned} \tag{3.3}$$

as $k \to \infty, k/n \to 0, n \to \infty$ and $k^{1/2} A(n/k) \to 0$, where A is from the second order condition, $\gamma_+ = \gamma \vee 0$, $\gamma_- = \gamma \wedge 0$, W is a Brownian motion, P_{γ_-} and R_{γ_-} are integrals involving W (for details, see Dietrich et al., 2002; de Haan and Ferreira, 2006), and $\hat{\gamma}_+$ and $\hat{\gamma}_-$ are consistent estimators of the corresponding parameters.

In practice, Dietrich et al. (2002) recommend replacing γ by its estimate. Therefore, based on (3.3), we could test

$$H_0 : F \in \mathcal{D}(G_\gamma), \gamma \in \mathbb{R} + \text{second order condition}$$

by first determining the corresponding quantile $Q_{1-\alpha,\hat{\gamma}}$ of the distribution $E_{\hat{\gamma}}$ and then comparing it with the value of $E_{k,n}$. If $E_{k,n} > Q_{1-\alpha,\hat{\gamma}}$ we reject H_0 with asymptotic type I error α and otherwise there is no evidence to reject H_0. Notice that this is a one-sided test of hypothesis, but a two-sided test could be fashioned similarly.

A drawback of this test is that we must include in H_0 the additional second

order condition, which is difficult to check in practice. While many common distributions satisfy the second order condition, including the normal, stable, Cauchy, log-Gamma, among others, the Pareto distribution is a notable example of a distribution which does not satisfy the second order condition. There are two other drawbacks of this test. First, it is based on the usual setting of acceptance-rejection regions, and thus it provides no measure of the strength of rejection of H_0. While this typically is addressed with the equivalent setting based on p-values, the limit distribution in (3.3) is analytically intractable and so are the p-values. Second, since the limit (3.3) depends on k, the conclusions of the test are also highly dependent on the choice of k.

Dietrich et al. (2002) observe that the limit distribution in (3.3) is simplified if $\gamma \geq 0$ since then $\gamma_- = 0$. Under the assumption that $F \in \mathcal{D}(G_\gamma), \gamma \geq 0$ and the second order condition, (3.3) becomes:

$$\begin{aligned}\tilde{E}_{k,n} =& k \int_0^1 \left(\frac{\ln Y_{n-[kt]:n} - \ln Y_{n-k:n}}{\hat{\gamma}_{k,n}} + \ln t \right)^2 t^2 dt \\ \xrightarrow{d} & \tilde{E} = \int_0^1 \left(t^{-1} W_t - W_1 + \ln t \int_0^1 (s^{-1} W_s - W_1) ds \right)^2 t^2 dt. \end{aligned} \tag{3.4}$$

Suppose $\tilde{E}_1, \ldots, \tilde{E}_N$ is a random sample of $\tilde{E}$, that we can obtain by simulation since the limit distribution in (3.4) is free of unknown parameters. Based on (3.4), construct a test for

$$H_0 : F \in \mathcal{D}(G_\gamma), \gamma \geq 0 + \text{second order condition},$$

as follows: Estimate a (one-sided) p-value $p(k) = \mathbb{P}(\tilde{E} > \tilde{E}_{k,n})$ as the relative frequency

$$\hat{p}(k) = \frac{1}{N} \sum_{j=1}^{N} 1_{\tilde{E}_j > \tilde{E}_{k,n}}.$$

If $\hat{p}(k) < \alpha$, then reject H_0 with an asymptotic type I error α; otherwise there is no evidence to reject H_0. With this method, the p-values give a mea-

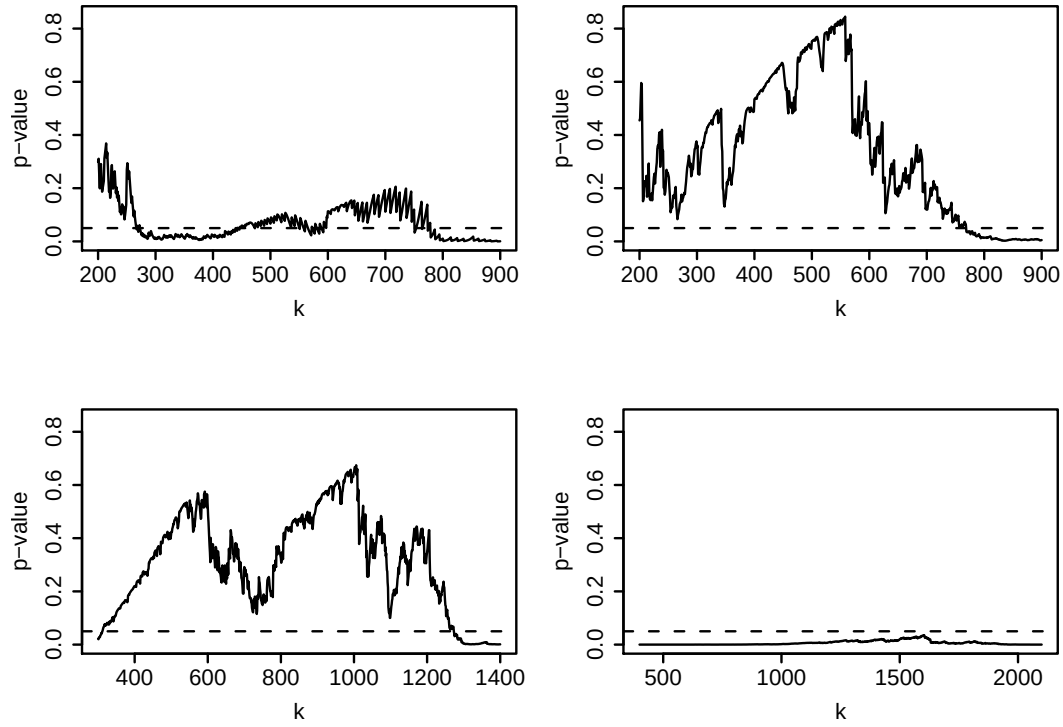

Figure 3.3: Plots of p-values as a funtion of k for the test of the extreme value condition for the marginal distributions of S, D, R. A horizontal dashed line is drawn at $\alpha = 0.05$. *Upper left*: Size in the 10th decile group. *Upper right*: Duration in the 10th decile group. *Lower left*: Rate in the 10th decile group. *Lower right*: Rate in the 4th decile group.

sure of the strength of rejection of H_0. Furthermore, we can check the stability of the conclusion of the test as a function of k by constructing the plot $\{(k, \hat{p}(k));\ k \text{ in an appropriate range}\}$. The range of values of k is chosen to accommodate for the limit (3.4), namely $k \to \infty, k/n \to 0, n \to \infty$. For example, Hüsler and Li (2006) found via simulations that the power of the test in Dietrich et al. (2002) appears to be high for k such that $k/n \approx 0.05$, at least for their various choices of F. To compute $\tilde{E}_{k,n}$, we use $\hat{\gamma}_{k,n}$ given by the consistent Hill estimator (Hill, 1975) or perhaps maximum likelihood if we suspect $\gamma = 0$.

We use this method with an asymptotic nominal type I error $\alpha = 0.05$. We found no evidence against $F_S, F_D \in \mathcal{D}(G_\gamma), \gamma \geq 0$ for all the 10 decile groups. Typical plots of the p-values $\hat{p}(k)$ for the variables S and D are those corresponding to the 10th decile group, shown in the upper left and upper right panels of Figure 3.3 respectively. Both plots exhibit $\hat{p}(k) > 0.05$ for a wide range of values of k. We also found that the plots $\{(k, \hat{p}(k))\}$ corresponding to the other decile groups show no evidence against $F_S, F_D \in \mathcal{D}(G_\gamma), \gamma \geq 0$. Coupled with the evidence from the QQ plots, we believe $\gamma > 0$.

Similarly, the lower left panel of Figure 3.3 exhibits the plot $\{(k, \hat{p}(k))\}$ corresponding to the distribution of R in the 10th decile group. Once again we found that $\hat{p}(k) > \alpha$ for a wide range of values of k, thus showing no evidence against $F_R \in \mathcal{D}(G_\gamma), \gamma \geq 0$. However, we did find evidence against $H_0 : F_R \in \mathcal{D}(G_\gamma), \gamma \geq 0 +$ *second order condition* for all the lower nine decile groups. A typical example of the plot $\{(k, \hat{p}(k))\}$ in these latter groups is exhibited in the lower right panel of Figure 3.3 for the 4th decile group, which shows that $\hat{p}(k)$ are significantly lower than 0.05 across a wide range of k values.

Therefore, for the lowest nine decile groups, we reject $H_0 : F_R \in \mathcal{D}(G_\gamma), \gamma \geq 0$+*second order condition*. One possible alternative is that $F_R \in \mathcal{D}(G_\gamma), \gamma < 0$, or equivalently, that x_{F_R}, the right endpoint of F_R is finite and $F_{(x_{F_R} - R)^{-1}} \in \mathcal{D}(G_{-1/\gamma})$ (de Haan and Ferreira, 2006; Resnick, 1987). By applying the above test to $H_0 : F_{(x_{F_R} - R)^{-1}} \in \mathcal{D}(G_\gamma), \gamma \geq 0$+*second order condition*, we rejected this new H_0 because the $\hat{p}(k) < 0.05$ for a wide range of values of k for the lower nine decile groups. Here we estimated x_{F_R} with $R_{n:n} + 1/n'$ for a high value of n'.

This last result leaves two possibilities. Either $F_R \notin \mathcal{D}(G_\gamma), \gamma \in \mathbb{R}$ or the

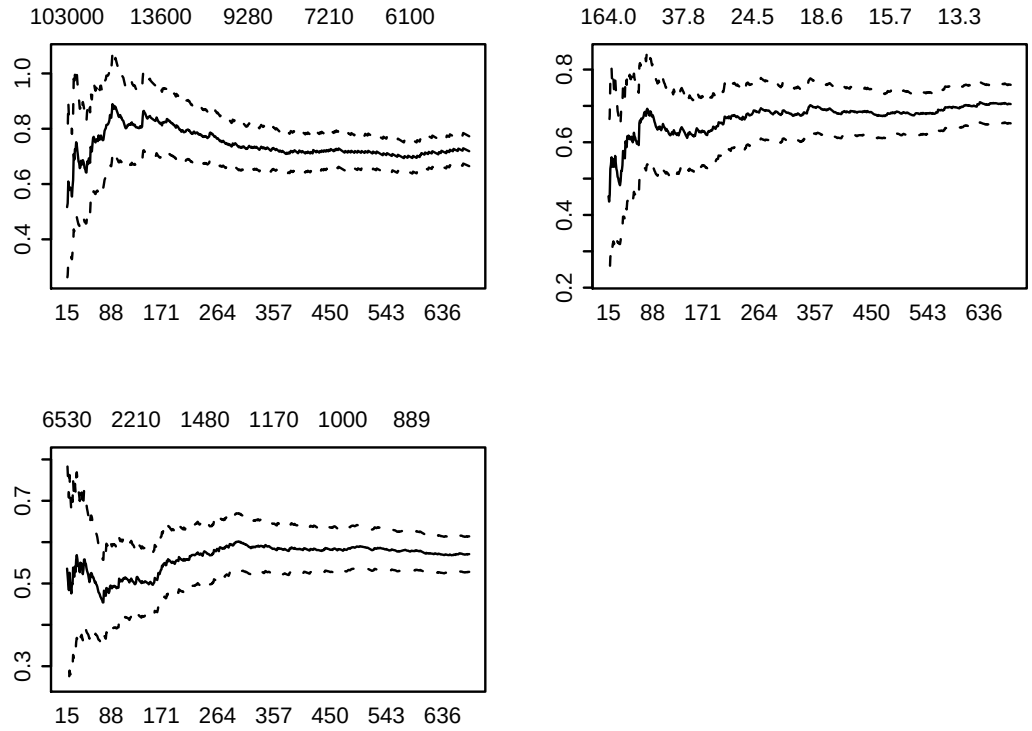

Figure 3.4: Hill plots for the shape parameter γ of the variables in the 10th decile group; dashed lines give 95% confidence bands. Values at the top of the plots give thresholds and the values on bottom indicate the number of upper order statistics. *Upper left*: Size. *Upper right*: Duration. *Lower left*: Rate.

additional second order condition does not apply for F_R (which does not rule out $F_R \in \mathcal{D}(G_\gamma), \gamma \geq 0$). Since the second order condition is difficult to check in practice, the constructed hypothesis test is somewhat vauge about F_R. In Section 3.5, we give more information about F_R.

3.3.2 Estimation

In Section 3.3.1 we showed that $F_S, F_D \in \mathcal{D}(G_\gamma), \gamma \geq 0$ for all the decile groups and $F_R \in \mathcal{D}(G_\gamma), \gamma \geq 0$ only for the 10th decile group. We now estimate the shape parameter γ for these distributions, using the consistent and asymptotically normal Hill estimator based on k largest order statistics, denoted as $\hat{\gamma}_{k,n}$ (Hill, 1975; Csörgő et al., 1985; Davis and Resnick, 1984; de Haan and Resnick, 1998).

Figure 3.4 exhibits Hill plots for the shape parameter γ of the distribution of S, D and R for the 10th decile group as a function of the number k of upper order statistics used (Hall, 1982; de Haan and Ferreira, 2006; Resnick, 2007; Geluk et al., 1997; de Haan and Resnick, 1998; Peng, 1998; de Haan and Peng, 1998; Mason and Turova, 1994). The three plots show stable regimes for γ around $k = 450$. We also found stability in the Hill plots for the shape parameter γ of the distribution of S, D in all the other decile groups.

Table 3.1 contains the Hill estimates of γ for our data set, along with estimates of the asymptotic standard error. A number k of upper order statistics was chosen individually for each variable and each decile group based on the corresponding Hill plots. For most decile groups, we used $k \approx 400$ $(k/n \approx 0.05)$, as suggested by the empirical study by Hüsler and Li (2006). Notice that the majority of the estimates are greater than 0.5, which implies that the corresponding distributions have infinite variances.

The Hill estimator requires $\gamma > 0$ but in Section 3.3.1 we only verified that $\gamma \geq 0$. Unlike the Hill estimator, the Pickands estimator (Pickands, 1975; Dekkers and de Haan, 1989) is valid for $\gamma \in \mathbb{R}$. However, the *Pickands plots*

Table 3.1: Summary of Hill estimates with asymptotic standard errors for the shape parameter of S, D and R.

Decile group	γ_S	s.e.	γ_D	s.e.	γ_R	s.e.
1	0.56	0.056	0.60	0.028		
2	0.55	0.061	0.47	0.023		
3	0.62	0.044	0.63	0.034		
4	0.62	0.036	0.62	0.029		
5	0.61	0.035	0.55	0.029		
6	0.69	0.040	0.55	0.028		
7	0.88	0.042	0.73	0.037		
8	0.77	0.045	0.71	0.033		
9	0.70	0.037	0.69	0.032		
10	0.73	0.034	0.68	0.032	0.58	0.027

proved to be very unstable for our data set and thus we relied on the Hill estimator which is close to the maximum likelihood estimator.

3.4 Dependence structure of (S, D, R) when the three variables have heavy tails

We now analyze the dependence structure of the triplet (S, D, R) across the 10 different decile groups. Since $S = DR$, at most two of the three components in (S, D, R) may be independent. This makes it reasonable to focus on the analysis of each pair of variables. We concentrate on the pairs in (S, D, R) with heavy tailed marginals and first focus on the dependence structure of (S, D) across the

10 deciles groups. We later study the dependence structure of both (R, S) and (R, D), but only in the 10th decile group. For the other decile groups, we found strong evidence suggesting R does not have heavy tails, and thus we leave this case for Section 3.5. Our finer segmentation into the deciles of $R^{\vee}$ reveals hidden features in an alpha/beta split, and therefore it is important to take into account the explicit level of $R^{\vee}$.

One way to assess the dependence structure is with sample cross-correlations. In heavy-tailed modeling, although the sample correlations may always be computed, there is no guarantee that the theoretical correlations exist. Recall Table 3.1 shows that most estimates of γ for S, D and R are greater than 0.5, and thus correlations do not exist in these instances. Moreover, correlation is a crude summary of dependence that is most informative between jointly normal variables. It does not separate dependence between large values and dependence between small values. In the context of data networks, the likelihood of various simultaneous large values of (S, D, R) may be important for understanding burstiness. For example, if large values of D are likely to occur simultaneously with large values of R, then we can expect a network that is prone to congestion. In this situation, a scatterplot $\{(D_i, R_i)\}$ would be mostly concentrated in the interior of the first quadrant of $\mathbb{R}^2$. On the other hand, if large values of one variable are not likely to occur with large values of the other one, the same scatterplot would be mostly concentrated on the axes.

Understanding network behavior requires a description of the extremal dependence of S, D and R and this extremal dependence is conveniently summarized by the spectral measure (de Haan and Resnick, 1977; de Haan and Ferreira, 2006; Resnick, 2007, 1987). We begin by discussing important concepts.

3.4.1 Bivariate regular variation and the spectral measure

Let $\mathbf{Z}$ be a random vector on $\mathbb{E} := [0,\infty]^2 \setminus \{(0,0)\}$, with distribution function F. The tail of F is *bivariate regularly varying* if there exist a function $b(t) \to \infty$ and a Radon measure ν on $\mathbb{E}$, such that

$$t\mathbb{P}\left[b(t)^{-1}\mathbf{Z} \in \cdot\right] \xrightarrow{v} \nu(\cdot), \tag{3.5}$$

vaguely in $\mathbb{E}$. This is the straightforward generalization of the univariate case as formulated in (1.2).

In terms of dependence structure of the components of $\mathbf{Z}$, it is often illuminating to consider the equivalent formulation of (3.5) that arises by transforming to polar coordinates. We define the *polar coordinate* transform of $\mathbf{Z} = (X,Y) \in \mathbb{E}$ by

$$(N,\Theta) = POLAR(\mathbf{Z}) := (\|\mathbf{Z}\|, \mathbf{Z}/\|\mathbf{Z}\|)\,, \tag{3.6}$$

where from this point on we use the L_1 norm given by $\|\mathbf{Z}\| = X + Y$.

Bivariate regular variation as formulated in (3.5) is equivalent to the existence of a function $b(t) \to \infty$ and a probability measure $\mathbb{S}$ on $\aleph_+ := \{\mathbf{z} \in \mathbb{E}; \|\mathbf{z}\| = 1\}$, such that

$$\mu_t(\cdot) := t\mathbb{P}\left[\left(b(t)^{-1}N, \Theta\right) \in \cdot\right] \xrightarrow{v} c\nu_\gamma \times \mathbb{S}(\cdot), \tag{3.7}$$

vaguely in $M_+((0,\infty] \times \aleph_+)$. Here $\nu_\gamma(r,\infty] = r^{-1/\gamma}$, $r > 0$, and $c > 0$ and, as usual, $M_+((0,\infty] \times \aleph_+)$ are the positive Radon measures on $(0,\infty] \times \aleph_+$. Since there is a natural bijection between $\aleph_+$ and $[0,1]$, namely $\mathbf{Z}/\|\mathbf{Z}\| \leftrightarrow X/\|\mathbf{Z}\|$, we can and will assume $\mathbb{S}$ is defined on $[0,1]$.

The probability measure $\mathbb{S}$, known as the *limit* or *spectral measure*, quantifies the asymptotic dependence structure of the bivariate random vector. Two cases

at opposite ends of the dependence spectrum (Coles, 2001; Resnick, 2007) are when (a) S concentrates on the two points $\{0, 1\}$, known as *asymptotic independence*, and (b) when S concentrates on $1/2$, known as *asymptotic full dependence*.

Since $\mathbb{S}$ could be any probability measure, there are infinitely many kinds of dependence structures between the two extreme cases discussed above. Therefore, we focus on the estimation of the spectral measure $\mathbb{S}$ as means of discerning the asymptotic dependence between two random variables with heavy tails.

3.4.2 Estimation of the spectral measure $\mathbb{S}$ by the antiranks method.

The definition (3.5) of bivariate regular variation requires scaling the two components of $\mathbf{Z} = (X, Y)$ by the same function $b(t)$. This implies that the distributions of both X and Y have the same shape parameters and that their distributions are tail equivalent (Resnick, 1971); this is the *standard regular variation* case. This is rarely encountered in practice and is not true for our variables of interest (S, D) (see Section 3.3.2). In order to estimate $\mathbb{S}$, one is required to transform to the standard case. A procedure that does not require estimation of the γs, yet achieves transformation to the standard case, thus allowing the estimation of $\mathbb{S}$, is the the antiranks method (Huang, 1992; de Haan and Ferreira, 2006; Resnick, 2007) which we now review.

For iid bivariate data $\{(X_i, Y_i), 1 \leq i \leq n\}$ from a distribution in a domain of attraction, define the marginal antiranks by

$$r_i^{(1)} = \sum_{l=1}^{n} 1_{[X_l \geq X_i]}, \quad r_i^{(2)} = \sum_{l=1}^{n} 1_{[Y_l \geq Y_i]},$$

and

- Transform the data $\{(X_i, Y_i), 1 \leq i \leq n\}$ using the antirank transform:

$$\{(X_i, Y_i); 1 \leq i \leq n\} \mapsto \{(k/r_i^{(1)}, k/r_i^{(2)}); 1 \leq i \leq n\}.$$

- Apply the polar coordinate transformation

$$POLAR\left(k/r_i^{(1)}, k/r_i^{(2)}\right) = (N_{i,k}, \Theta_{i,k}).$$

- Estimate $\mathbb{S}$ with (Resnick, 2007; de Haan and Resnick, 1993)

$$\hat{\mathbb{S}}_{k,n}(\cdot) = \frac{\sum_{i=1}^{n} \epsilon_{(N_{i,k},\Theta_{i,k})}((1,\infty] \times \cdot)}{\sum_{i=1}^{n} \epsilon_{N_{i,k}}((1,\infty])} \Rightarrow \mathbb{S}(\cdot). \tag{3.8}$$

The interpretation of (3.8) is that the empirical probability measure of those Θs whose radius N is greater than 1 consistently approximates $\mathbb{S}$. Hence, we get an estimate of $\mathbb{S}$ by fitting a distribution to the points $\{\Theta_{i,k};\ N_{i,k} > 1\}$, for a suitable k (see Section 3.4.3). Though we do not know that $\mathbb{S}$ has a density, often a density estimate is more striking than a distribution function estimate. For example, a mode in the density at $1/2$ reveals a tendency towards asymptotic dependence, but modes in the density at 0 and 1 exhibit a tendency towards asymptotic independence.

3.4.3 Parametric estimation of the spectral density of (S, D)

Using the antiranks method, we transform the points $\{(S_i, D_i)\}$ for each decile group separately. Figure 3.5 shows 10 histograms of the transformed points $\{\Theta_{i,k}; N_{i,k} > 1\}$. The histograms suggest decreasing strength of dependence between S and D as $R^{\vee}$ increases, since mass is increasingly distributed away

from the midpoint of $[0, 1]$. Certainly asymptotic independence does not hold in any decile group.

To investigate this suggestion, we fit a parametric family to the spectral density. The histograms in Figure 3.5 are reasonably symmetric for each decile group, suggesting that the logistic family (Coles, 2001) may be an appropriate parametric model. The logistic family

$$h_\psi(t) = \frac{1}{2}\left(\frac{1}{\psi} - 1\right) t^{-1-\frac{1}{\psi}}(1-t)^{-1-\frac{1}{\psi}}[t^{-\frac{1}{\psi}} + (1-t)^{-\frac{1}{\psi}}]^{\psi-2}, \quad 0 \le t \le 1, \quad (3.9)$$

is a symmetric model with a single parameter $\psi \in (0, 1)$. For $\psi < 0.5$, h is unimodal, whereas for increasingly large values of $\psi > 0.5$, the density places greater mass towards the ends of the interval $[0, 1]$. In fact, asymptotic independence holds as $\psi \to 1$, and perfect dependence is obtained as $\psi \to 0$. This allows us to quantify the effect of $R^\vee$ on the dependence between S and D.

We first fit the model (3.9) to the data $\{\Theta_{i,k}; N_{i,k} > 1\}$ within each $R^\vee$ decile group by maximum likelihood estimation. The log-likelihood function of ψ based on $t_1, \ldots, t_n$ is

$$\begin{aligned} l(\psi) = &\sum_{i=1}^{n} \ln\left(\frac{1}{\psi} - 1\right) - \sum_{i=1}^{n}\left(1 + \frac{1}{\psi}\right)\ln(t_i(1-t_i)) \\ &+ \sum_{i=1}^{n}(\psi - 2)\ln(t_i^{-1/\psi} + (1-t_i)^{-1/\psi}), \end{aligned} \quad (3.10)$$

which we maximize numerically for $0 \le \psi \le 1$. By considering ψ as a function of k, we choose a value of k around which the estimate of ψ looks stable. Figure 3.5 shows that the logistic estimates of the spectral density are in close agreement to the histogram of the points. On top of each plot, we indicate the maximum likelihood estimates of ψ and the choice of k in the corresponding decile group. The estimates of ψ confirm a decline in dependence between S and D as the decile group increases, as measured by increasing estimates of ψ.

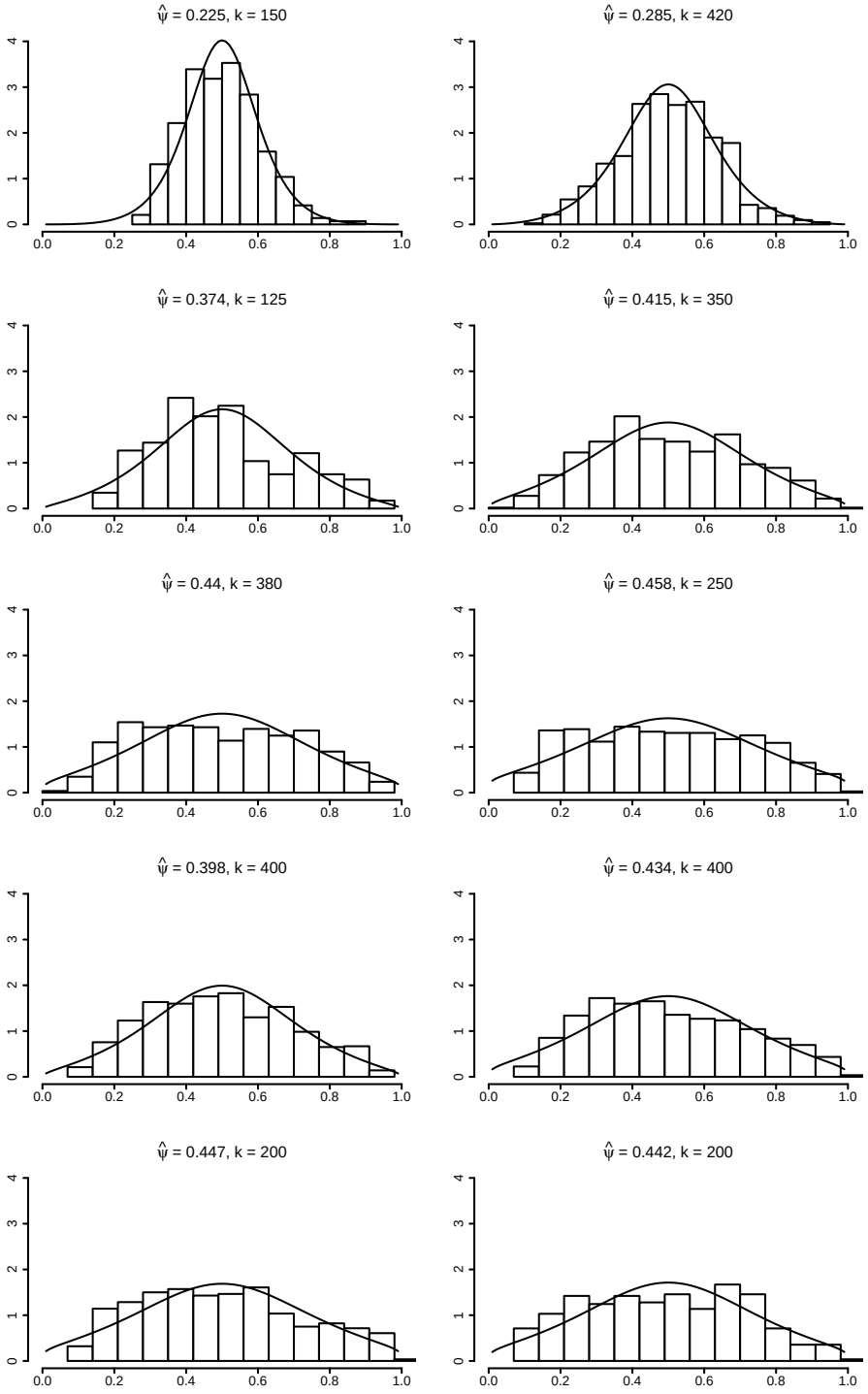

Figure 3.5: Logistic MLE estimates of the spectral density of (S, D) superimposed on the histograms of the points $\{\Theta_{i,k}; N_{i,k} > 1\}$, starting with the 1st decile group from the upper left and going left to right by row.

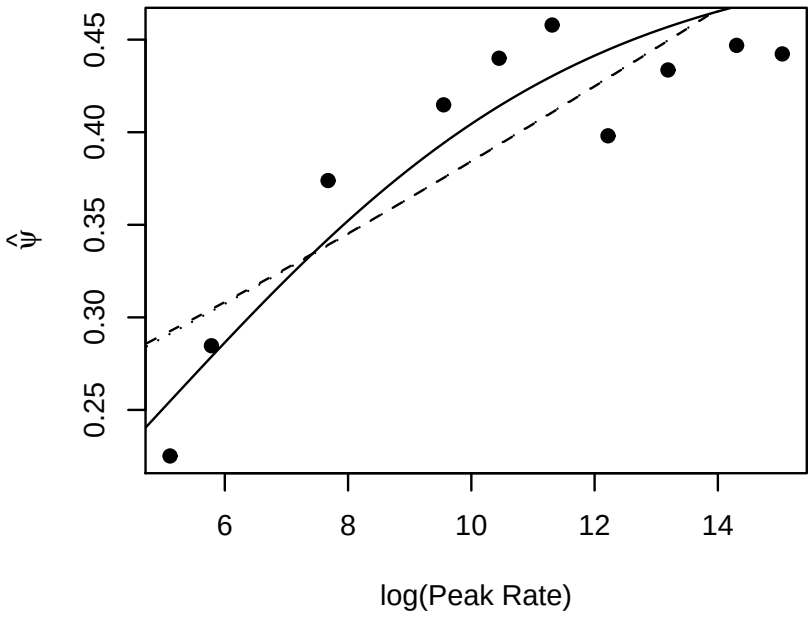

Figure 3.6: Parameter ψ as a function of $\ln(R^{\vee})$ and three linear models of the form (3.11) superimposed: (solid curve) link function (3.12), (dashed line) logit link, (dotted line) probit link. The logit and probit links are almost indistinguishable in the range of the data.

We now study the form of this decline by fitting a global trend model simultaneously to all the peak rate decile groups, using the same data (antirank transformed, polar coordinate transformed, thresholded) employed for the separate analyses. In this joint study, the parameter ψ in (3.10) is a function of $R^{\vee}$ as follows:

$$g^{-1}(\psi) = \beta_0 + \beta_1 \ln(R^{\vee}), \tag{3.11}$$

where g is a link function. The used of $\ln(R^{\vee})$ instead of $R^{\vee}$ is a common technique in linear models to improve fit. Since $\psi \in (0, 1)$, natural choices of g are the logit and the probit functions. However, as shown in Figure 3.6, the link

function

$$g(x) = \frac{0.5}{1 + e^{-x}} \tag{3.12}$$

is more adequate than the usual logit or probit links. Our link is very similar to the logit link, but confines possible values of ψ to the interval $(0, 0.5)$ and is suggested by the fact that in Figure 3.5 the histograms of the points $\{\Theta_{i,k}; N_{i,k} > 1\}$ put all mass around an apparent mode at 0.5. This behavior corresponds to $\psi < 0.5$.

Figure 3.6 exhibits in various ways the logistic parameter ψ as a function of peak rate using (3.11). First, we plot the points $\mathcal{P} = \{(med^{(i)}, \hat{\psi}^{(i)}); 1 \leq i \leq 10\}$, where $med^{(i)}$ is the median of the $\ln R^{\vee}$ variable for sessions in the ith decile group and $\hat{\psi}^{(i)}$ is the maximum likelihood estimated logistic parameter in the ith decile group. In Figure 3.6, we superimpose on $\mathcal{P}$ the estimated (3.11) using the link function (3.12), showing that the goodness of fit of the model (3.11) is quite reasonable.

To assess the effect of $R^{\vee}$ on the dependence structure of (S, D), we focus on $\hat{\beta}_1$. Observe that (3.10) gives the log-likelihood of the model for independent observations. Since $\{\Theta_{i,k}; N_{i,k} > 1\}$ is not an independent sample due to the antirank transform, the classical maximum likelihood theory is not strictly applicable. Hence, to quickly compute the standard error of $\hat{\beta}_1$ we bootstrap the whole model. However, several authors have shown in the context of heavy-tailed phenomena that if the original sample is of size n, then the bootstrap sample size m should be of smaller order for asymptotics to work as desired (Athreya, 1987; Deheuvels et al., 1993; Giné and Zinn, 1989; Hall, 1990; Resnick, 2007). In connection with the estimation of the spectral measure, the bootstrap procedure works as long as $m \to \infty$, $m/n \to 0$ and $n \to \infty$. Therefore, a boot-

strap procedure to estimate the standard error of $\hat{\beta}_1$ is constructed as follows:

(i) From the original sample $\{(S_i, D_i, R_i^{\vee}); 1 \leq i \leq 44136\}$, a bootstrap sample $\{(S_i^*, D_i^*, R_i^{\vee *}); 1 \leq i \leq 10000\}$ is obtained with the bootstrap sample size of smaller order than the original sample size. Our choice of $m = 10000$ results from the need to have sufficient data for estimation. (Choosing the bootstrap sample size is as difficult as choosing the threshold k in, say, Hill estimation.)

(ii) Split the bootstrap sample $\{(S_i^*, D_i^*, R_i^{\vee *}); 1 \leq i \leq 10000\}$ into 10 groups according to the quantiles of $R_i^{\vee *}$.

(iii) Within each bootstrap decile group, transform the data $\{(S_i^*, D_i^*); 1 \leq i \leq 1000\}$ using the antirank transform and then transform to polar coordinates to obtain $\{\Theta_{i,k}^*; N_{i,k}^* > 1\}$. Here, for each bootstrap decile group we use the same value of k that is used in the original estimation. These values are shown in Figure 3.5.

(iv) Fit the global linear trend simultaneously to all the bootstrap decile groups, by maximizing (3.10) with ψ as a function of $R^{\vee *}$ as in (3.11) and (3.12). Hence, we obtain a bootstrap replication $\hat{\beta}_{1,b}^*$.

(v) Repeat steps (i)-(iv) $B = 1000$ times and estimate the standard error of $\hat{\beta}_1$ by the sample standard deviation of the B replications

$$\widehat{se}(\hat{\beta}_1) = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} [\hat{\beta}_{1,b}^* - \hat{\beta}_1^*]^2 \right\}^{1/2}, \tag{3.13}$$

where $\hat{\beta}_1^* = \sum_{b=1}^{B} \hat{\beta}_{1,b}^*/B$.

Table 3.2 summarizes the estimated parameters of the linear model for ψ and their standard errors. Our model assesses dependence through the value

Table 3.2: Summary of estimated linear model given by (3.11) and (3.12).

	Estimated parameter	Bootstrap standard errors
$\hat{\beta}_0$	-1.432	0.219
$\hat{\beta}_1$	0.288	0.127

of ψ and from our results, we conclude that $R^{\vee}$ exerts a significant effect on the dependence structure of (S, D), since $\hat{\beta}_1$ is significantly different from 0. Jointly, (3.11) and (3.12) provide an adequate description of the behavior of ψ across the decile groups.

3.4.4 Parametric estimation of the spectral density of (R, S) and (R, D)

We now transform the points $\{(R_i, S_i)\}$ and the points $\{(R_i, D_i)\}$ in the 10th decile group using the previously described antirank transform. Figures 3.7(a) and 3.7(b) exhibit histograms of the transformed points $\{\Theta_{i,k}; N_{i,k} > 1\}$ corresponding to the pairs (R, S) and (R, D), respectively. Both histograms look reasonably symmetric, and thus the modeling is done via the logistic family (3.9).

Figure 3.7 shows that the fitted logistic models are in close agreement with the empirical distribution of the points $\{\Theta_{i,k}; N_{i,k} > 1\}$. Notice that the parameter ψ of the logistic density corresponding to the pair (R, D) is closer to 1 than the parameter ψ of the density corresponding to the pair (R, S). This suggests that for the group of sessions with the highest values of $R^{\vee}$, the scheme RD (in

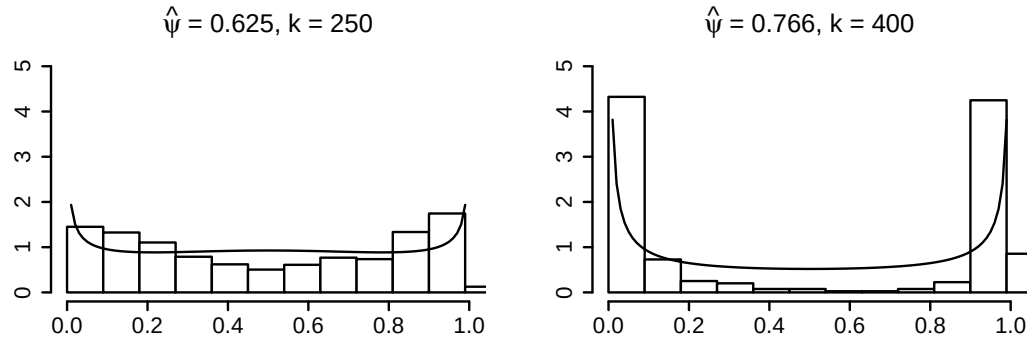

Figure 3.7: Logistic estimates in the 10th decile group superimposed on the histograms of the points $\{\Theta_{i,k}; N_{i,k} > 1\}$. *Left:* Spectral density of (R, S). *Right:* Spectral density of (R, D).

which R and D are independent, at least asymptotically), is more adequate than the scheme RS (in which R and S are independent, at least asymptotically). This conclusion is exactly the opposite to Sarvotham et al. (2005)'s, since they recommend using the scheme RS for the group with the highest peak rates (that is, their alpha group).

The fact that for the sessions with the highest values of peak rate $R^{\vee}$, we have (R, D) close to asymptotically independent may have the following interpretation. Users with high bandwidth pay little or no attention to the duration of their downloads; this is expected because such users know that probably their lines are capable of downloading any file, no matter how long it takes.

3.5 Dependence structure of (S, D, R) when R does not have heavy tails

We now investigate the dependence structure of (R, S) and (R, D) in the first nine decile groups, that is, those with values of $R^{\vee}$ in the decile ranges $(10(g-1)\%, 10g\%]$, $g = 1, \ldots, 9$. For these groups, there is evidence that the distribution of R is not heavy tailed. Moreover, the diagnostics in Section 3.3.1 suggest that $R \notin \mathcal{D}(G_\gamma)$ for any $\gamma \in \mathbb{R}$. However, the other variables S and D have heavy tails in these decile groups, and we can make use of the conditional extreme value model (Heffernan and Tawn, 2004; Heffernan and Resnick, 2007; Das and Resnick, 2011a,b) to study the dependence structure of the pairs (R, S) and (R, D).

3.5.1 The conditional extreme value model

Classical bivariate extreme value theory assumes that both variables are in some maximal domain of attraction. When one variable is in a domain of attraction, but the other is not, the conditional extreme value (CEV) model is a candidate model.

Let $\mathbf{Z} = (X, Y) \in \mathbb{E} = [0.\infty]^2 \setminus \{(0, 0)\}$ and let $\bar{\mathbb{E}}^{(\gamma)}$ be the right closure of $\mathbb{E}^{(\gamma)} = \{y \in \mathbb{R} : 1 + \gamma y > 0\}$. The CEV model assumes that $F_Y \in \mathcal{D}(G_\gamma), \gamma \in \mathbb{R}$, with normalizing sequences $a(t) > 0$ and $b(t)$ as in (1.4). In addition, the CEV model assumes that there exist functions $\alpha(t) > 0$, $\beta(t) \in \mathbb{R}$ and a non-null Radon measure μ on the Borel subsets of $[-\infty, \infty] \times \bar{\mathbb{E}}^{(\gamma)}$ such that the following conditions hold for any $y \in \mathbb{E}^{(\gamma)}$:

(i) For $\mu-$continuity points (x, y):

$$t\mathbb{P}\left(\frac{X-\beta(t)}{\alpha(t)} \leq x, \frac{Y-b(t)}{a(t)} > y\right) \to \mu([-\infty, x] \times (y, \infty]), \quad t \to \infty. \tag{3.14}$$

(ii) $\mu([-\infty, x] \times (y, \infty])$ is not a degenerate distribution in x.

(iii) $\mu([-\infty, x] \times (y, \infty]) < \infty$.

(iv) $H(x) := \mu([-\infty, x] \times (0, \infty])$ is a probability distribution.

3.5.2 Method for verifying the CEV model

Das and Resnick (2011b) recently developed a method for checking the adequateness of the CEV model. Suppose $\{(X_i, Y_i); 1 \leq i \leq n\}$ are iid from the CEV model. Define:

- $Y_{(1)} \geq \ldots, Y_{(n)}$: The upper-order statistics of $Y_1, \ldots, Y_n$.
- $X_i^*, 1 \leq i \leq n$: The X-variable corresponding to $Y_{(i)}$, also called the *concomitant* of $Y_{(i)}$.
- $r_{i,k}^* = \sum_{l=1}^{k} 1_{[X_l^* \leq X_i^*]}$: The rank of X_i^* among $X_1^*, \ldots, X_k^*$.

The *Hillish statistic* of $\{(X_i, Y_i); 1 \leq i \leq n\}$ is defined as

$$\text{Hillish}_{k,n} := \frac{1}{k} \sum_{j=1}^{k} \ln \frac{k}{r_{i,k}^*} \ln \frac{k}{j}.$$

Under $H_0 : \{(X_i, Y_i); 1 \leq i \leq n\}$ *are iid from a CEV model,* Das and Resnick (2011b) proved that as $k \to \infty, k/n \to 0$, and $n \to \infty$:

$$\text{Hillish}_{k,n} \xrightarrow{P} I_{\mu,H}, \tag{3.15}$$

where $I_{\mu,H}$ is a constant that depends on μ and H defined in Section 3.5.1.

Like the Hill estimator, the Hillish statistic depends on the number k, so we make a *Hillish plot* $\{(k, \text{Hillish}_{k,n}); k \geq 1\}$ and observe whether the plot has a stable regime. If that is the case, we conclude that the CEV model is adequate for (X, Y).

3.5.3 Verifying the CEV model for (R, S)

The CEV model appears as a candidate model for (R, S) or (R, D) for any one of the lowest 9 decile groups, since for these groups R does not appear to be in a domain of attraction, while both S and D have heavy tails. We found that the CEV model is adequate for (R, S) within each of the lowest nine $R^{\vee}$-decile groups. Here we present our results.

Figure 3.8 shows Hillish plots for checking the CEV model for (R, S) in the lowest nine decile groups. Apart from the second and third decile groups, all the plots look exceptionally stable. Although the plots do not look as good for the second and third decile groups, we still find a stable regime about the $k = 800$ upper order statistic, which supports the CEV model for these groups as well. This emphasizes that more detailed structure exists for the beta group of Sarvotham et al. (2005) and that further segmentation reveals more information.

Moreover, observe that the limit constant $I_{\mu,H}$ varies with the decile group. In effect, $I_{\mu,H}$ decreases as $R^{\vee}$ goes up. Since $I_{\mu,H}$ depends on the limit H in (3.14), this suggests that the conditional distribution of R given S varies with the decile group and that the dependence structure of (S, D, R) depends on the explicit level of $R^{\vee}$.

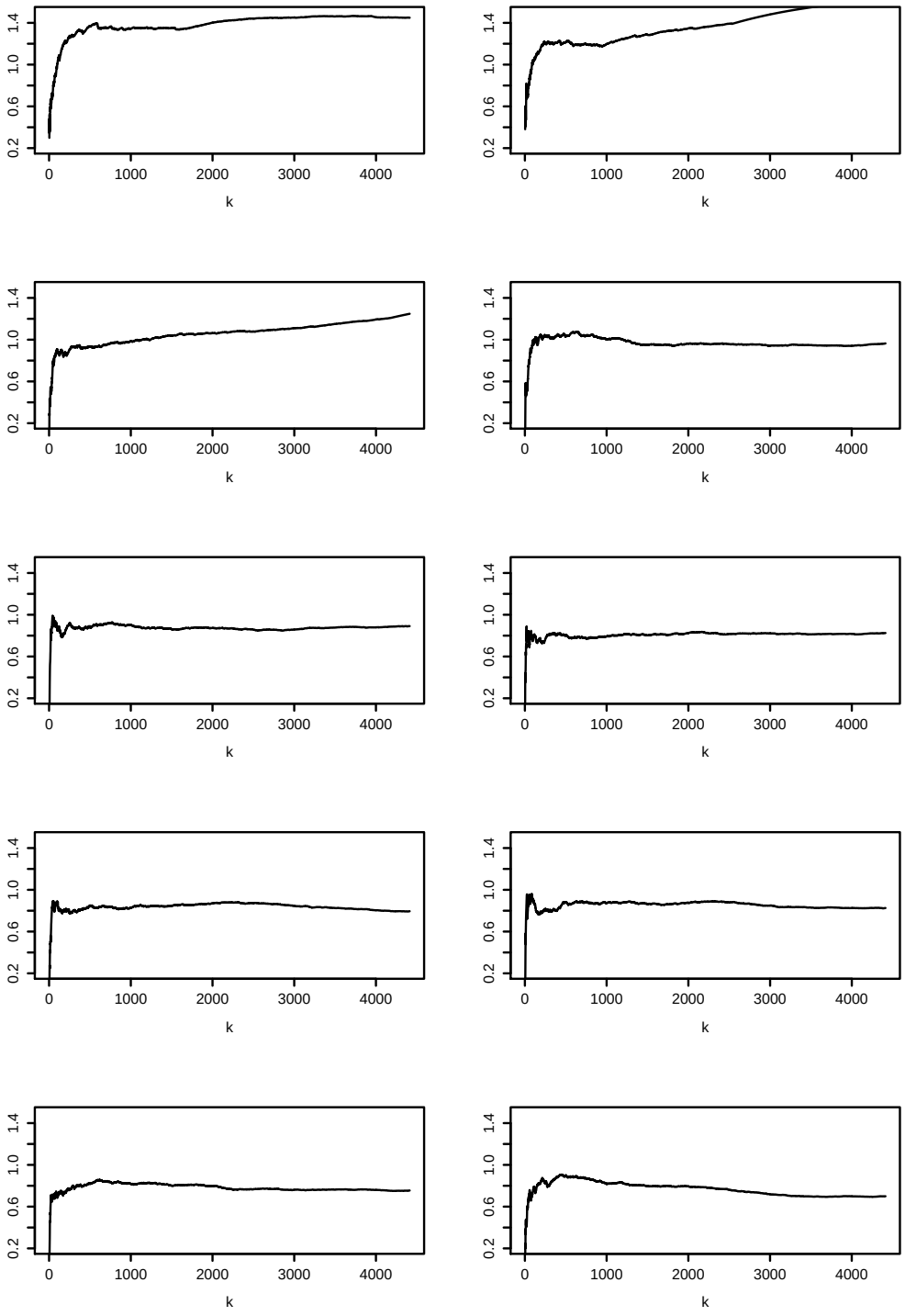

Figure 3.8: Hillish statistic of (R, S), starting with the 1st decile group from the upper left and going by row.

In addition, the Hillish plots reject the CEV model for the pair (R, D) in all the decile groups. We have not displayed these plots.

3.6 The Poisson property

There is considerable evidence against the Poisson model as the generating mechanism for network traffic at the packet-level (Paxson and Floyd, 1995; Willinger et al., 1997; Willinger and Paxson, 1998; Hohn et al., 2003). However, the classical explanation of Poisson arrival times, namely human activity generating independent activity, each with small probability of occurrence, is still applicable to network traffic aggregated to higher levels. A significant example is provided by Park et al. (2006), who show that "navigation bursts" in the server occur according to the Poisson model.

For our data, we found that the Poisson model does not appear to activate overall network traffic, but it does initiate user sessions for any given group of sessions whose peak rate $R^{\vee}$ is in a fixed inter-decile range. This allows for straightforward simulation within each decile group via a homogeneous Poisson process. This result depends on segmenting using our definition of peak rate and fails to hold when segmenting using either R, R_δ or I_δ.

Recall we split the sessions into 10 groups according to the deciles of $R^{\vee}$. For any given decile group, suppose that Γ_i are the starting times of the user sessions in increasing order; if necessary, we relabel sessions within the group. Let $\Delta_i = \Gamma_{i+1} - \Gamma_i$ be the session interarrival times. A homogeneous Poisson process is characterized by $\{\Delta_i\}$ being iid with the exponential $\exp(\lambda)$ as the common distribution function, for some parameter $\lambda > 0$.

3.6.1 Checking the exponential distribution for interarrival times

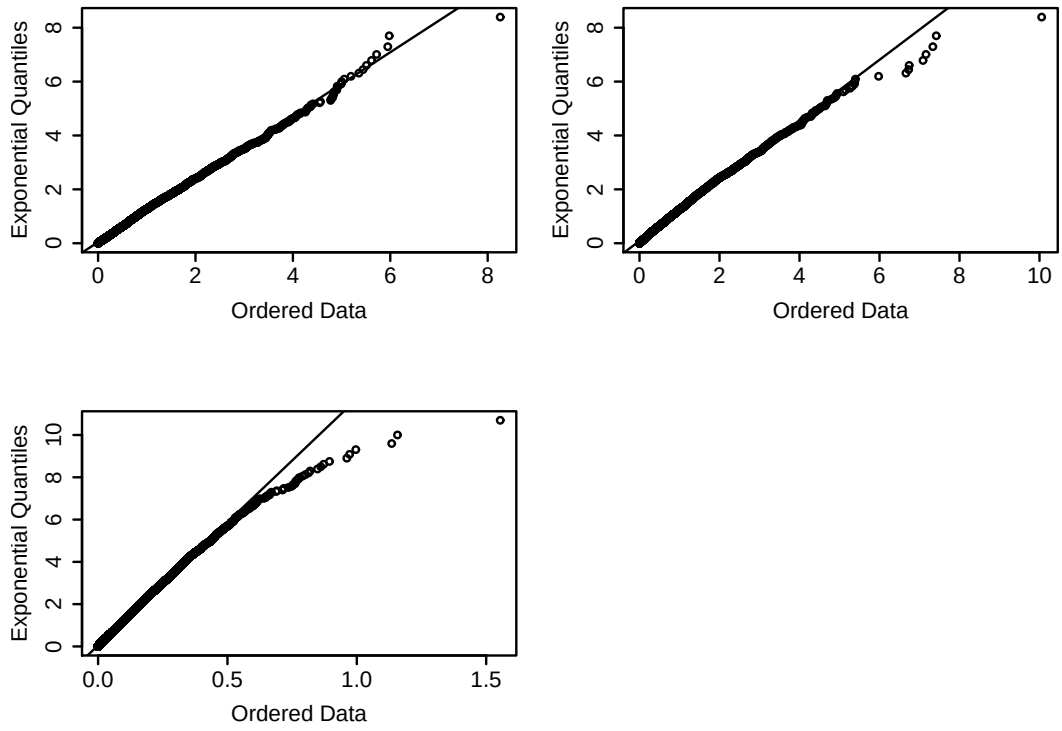

Figure 3.9: Exponential QQ plots of the interarrival times of sessions. *Upper left*: 4th decile group. *Upper right*: 10th decile group. *Lower left*: Overall traffic.

We verified that $\{\Delta_i\}$ may be accurately modeled as exponential random variables within each $R^\vee$ decile group. As examples, the upper left and upper right panels of Figure 3.9 exhibit exponential QQ plots of $\{\Delta_i\}$ for the 4th and 10th decile groups, respectively, which compare the quantiles of the empirical and theorical distributions. It is striking how well a straight line trend is shown, and this result replicates across all the decile groups. However, when all the sessions are put together in a single population, the session interarrival times

have right tails noticeably heavier than exponential, as shown in the lower left panel of Figure 3.9.

Interestingly, we found that the interarrival times within each decile group are not exponentially distributed if we segment sessions by deciles of either of the two previous predictors of burstiness I_δ and R_δ.

3.6.2 Independence of interarrival times

Within each decile group, can we use the independence model for $\{\Delta_i\}$? We investigated this question using the *sample autocorrelation function (acf)*. From Section 3.6.1, we know $\{\Delta_i\}$ can be modeled as an exponential random sample, and thus we can safely assume that the variances of Δ_i are finite. Therefore, the standard L_2 theory applies and Bartlett's formula from classical time series analysis (Brockwell and Davis, 1991) provides asymptotic normality of the sample acf under the null hypothesis of independence.

The left and right panels of Figure 3.10 exhibit sample acf plots for $\{\Delta_i\}$ for the 4th and 10th decile groups. In each figure, we plot the confidence bounds for an $\alpha = 0.05$. We counted 178 and 141 "spikes" exceeding the bounds, which is less than 5% of the total of 4414. In general, we found that less than 5% of the spikes lie outside the bounds for all the decile groups. Based on the sample acf, there is no evidence against the independence of $\{\Delta_i\}$ within each decile.

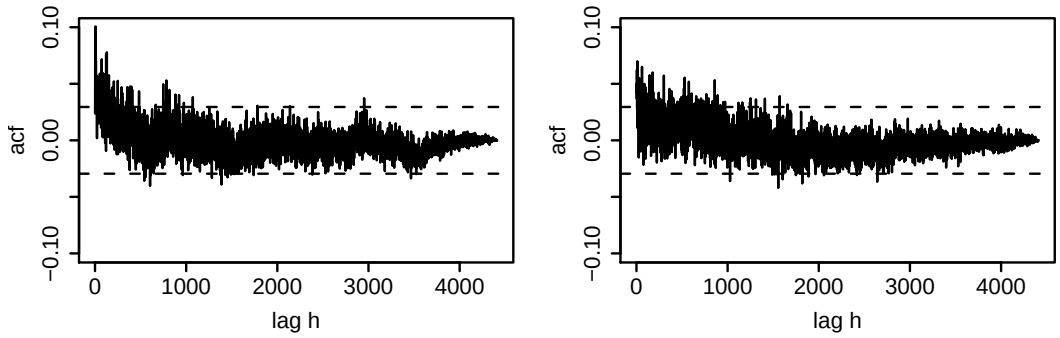

Figure 3.10: Sample autocorrelation functions of Δ_i. *Left:* 4th decile group. *Right:* 10th decile group.

3.7 Final remarks and conclusions

For the purposes of illustration, we have presented our analysis on a data set publicly available as of June 2011 through the Réseaux IP Européens (French for "European IP Networks") Network Coordination Centre's (RIPE NCC) data repository at `http://labs.ripe.net/datarepository`. The particular data file chosen for the analysis is "19991207-125019", which can be found within a collection of network data traces dubbed *Auckland II* recorded in 1999. We have successfully tested our analyses and proposed models given by (3.9) and (3.11) for other data files in the collection *Auckland II*, as well as a more recent collection dubbed *Auckland VIII* recorded in 2003. Unfortunately, the RIPE NCC website is not likely to live forever, as funding of such repositories come and go. For example, the same data set was previously available at the National Laboratory for Applied Networking Research, which was shutdown in May 2009; later, the Waikato Internet Traffic Storage took over some of the data sets, but delegated the responsibility to RIPE NCC soon after.

The reason for our choice of the logistic family and the linear trend is be-

cause they allow for a simple description of the dependence structure of (S, D) via the logistic parameter. As depicted in Figures 3.5 and 3.6, the proposed logistic model defined by (3.9) and (3.11), does a sound job of explaining the dependence structure of (S, D) as a function of the explicit level of the peak rate $R^{\vee}$.

Our findings yield an accurate simulation method for generating network sessions from the asymptotic model as follows:

1. Bootstrap from the empirical distribution of $R^{\vee}$ and split the bootstrap sample into, say, 10 groups according to the empirical deciles.
2. Conditionally on the decile group, simulate the starting times Γ of the sessions according to a homogeneous Poisson process. The Poisson rate depends on the decile group so from the original data set estimate the Poisson rates for each decile group and use them here.
3. For each synthetic $R^{\vee}$, compute ψ using (3.11) with the estimated values $\hat{\beta}_0$, $\hat{\beta}_1$. Use ψ to simulate an "angle" Θ from the logistic density h_{ψ}.
4. Simulate the radial component N as an independent heavy tailed random variable, for instance the Pareto (Resnick, 2007; de Haan and Resnick, 1993).
5. Transform (N, Θ) to Cartesian coordinates in order to get (S, D) in the standard case coordinate system.
6. Finally, power up (S, D) to a different exponent to adjust for possibly different marginal tail behavior.

We are considering details of a software procedure to implement this simulation suggestion.

We have shown evidence for the two following models:

- The classical extreme value theory for the pair (S, D), in which both components are heavy-tailed.
- The conditional extreme value (CEV) model for the pair (R, S), in which only one component, namely S, is heavy-tailed.

Given the fact that $R = S/D$, the question remains open about what conditions on the CEV model for (R, S) imply the classical model for (S, D), and vice versa.

Our analyses need to be extended to other segmentation schemes. For instance, heterogeneous traffic comprising different types of applications undoubtedly behaves differently from more homogeneous traffic, a fact used to justify the modeling in D'Auria and Resnick (2008). In the next chapter, we pursue extreme value analysis of traffic segmented by application type.

CHAPTER 4

MODELING NETWORK APPLICATION ACTIVITY

4.1 Overview

Network researchers deem knowledge of applications as essential for guaranteeing the quality of service offered to end-users, preventing congestion and bottlenecks, and identifying malicious traffic.

D'Auria and Resnick (2006, 2008) suggested that the tail behavior, dependence structure, and distributions of key session features may depend on the statistical characteristics of each network application. For instance, session sizes clearly differ according to the generating application (e.g. a session devoted to email would consist of a much smaller number of bytes than one devoted to streaming a movie). The same happens with session durations and rates.

However, statistical characterization of network applications poses two big challenges:

- How do we identify applications from packet headers? Packet headers contain information on network *ports*, which are software-based labels of connection endpoints that allow network applications to share hardware resources without interfering with each other. Well-known ports are associated with applications like web or email, but newer applications such as peer-to-peer file sharing, streaming media or network gaming use not only one, but a range of ports, or dynamically-allocated ones.
- How do we estimate tail behavior under censoring? Measurement studies

collect data over a fixed interval but sessions may start or end outside the measurement interval; such sessions experience a form of censoring.

In order to identify applications from network packet headers, many complex methodologies have been developed. Nevertheless, all of them belong to at least one of the following classes:

- Analysis of the client-server architecture of applications. For some applications, direct communication between clients never occurs, while for some others, each node can act both as a client and a server simultaneously. (See e.g. Kim et al., 2003; Karagiannis et al., 2004; Kim et al., 2005; Wang and Liu, 2007; Shane et al., 2007).
- Examination of some or all packets payload (Wang and Liu, 2007; Dharmapurikar et al., 2004). Publicly available data traces as well as most proprietary data sets do not contain information about packet content. Having payload information would certainly elicit concern over privacy and increase storage and processing overhead. In addition, usual methods of anonymization and encryption make examination of packet content difficult.
- Machine-learning analysis of session features (Hernandez-Campos et al., 2005; Moore and Zuev, 2005; Crotti et al., 2007; Cao et al., 2008; Maiolini et al., 2009). These include various supervised, unsupervised and semi-supervised methods for constructing classifiers, but many of these rules are obscure and nonintuitive.

It is not our goal to propose a new identification methodology or to compare the existent ones. Rather, the principal motivation of this chapter is to answer

whether we can do statistical analysis of network applications along the same lines of Chapter 3. For this purpose, we have relied on the traditional segmentation by ports, and focus on the statistical challenges such as the aforementioned censoring.

Section 4.2.1 reviews the identification of applications using network ports, and discusses its flaws. Section 4.2.2 describes censoring types ocurring in data networks.

In Section 4.3 we study the marginal distributions of size and duration, checking whether they belong to a domain of attraction $\mathcal{D}(G_\gamma), \gamma > 0$. Although Hill plots of the shape parameter γ_D of the session durations look stable, QQ plots signal the presence of censoring that may bias these estimates. So we use a method for estimating the marginal distribution of session duration proposed by Beirlant and Guillou (2001), designed censored observations. We show that durations have heavier tails for applications such as flash than for web, secure web, and email. Beirlant and Guillou (2001)'s estimator does not account for all types of censoring, and it is not directly applicable to session size and thus thus we also discuss a maximum likelihood estimator based on POT that accounts for all types of censoring and use it to estimate the shape parameter of both size and duration.

For each application, we verify in Section 4.4 that sessions do not arrive according to a Poisson process. However, there are periods in which the Poisson assumption does hold; these periods are terminated by network disruptions after which sessions from a given application take longer to arrive. Then the Poisson assumption resumes.

In Section 4.5 we give some remarks, including comments on identification of ports using clustering techniques on (S, D).

4.2 Basic concepts

4.2.1 Identification of applications using well-known ports

The Internet Assigned Numbers Authority (IANA) is responsible for the allocation of globally unique names and numbers that identify machines, servers and various network devices. IANA publishes these identifiers in memorandums known as Request For Comments (RFC). RFCs also describe methods, behaviors, research, or innovations applicable to the working of the Internet, such as the regulation of the transport layer and the registration of network ports.

The transport layer ensures the reliable arrival of packets, and provides error checking mechanisms and data flow controls. The two most known transport protocols are the Transport Control and the User Datagram protocols (TCP and UDP, respectively). Different applications are associated with different transport protocols. For example, web, mail, FTP, SSH, and some peer-to-peer file sharing typically use TCP, whereas UDP is used by streaming, VoIP and network gaming.

Network ports sit below the transport layer. These are part of a software system that allows computers to simultaneously handle multiple networking tasks by dividing network traffic into a series of individual feeds so information and services stay separate. Network *ports* are divided into three non-overlapping

categories: the *well-known* ports, the *registered* ports, and the dynamic or *private* ports. Common applications such as web or email transmit information through a process known as *listening*. Consider the following example: When you want to read say, the New York Times website (stored in network node B), you type its web address in your browser (node A), which asks node B for permission to access the website content. If everything goes well, node B approves your access and it asks node A whether it is ready to receive the New York Times website; node B knows A will be listening to this question through port 80, since it is the well-known port through which website requests are delivered. Finally, node A sends an acknowledgement (that it is ready) through one of the node B's private ports, and B starts sending the New York Times website content. In network terminology, node B is the *source* (of information) and node A is the *destination*. If there are two or more transmissions of the same type of service (say HTTP) between a source and a destination, the pair *(source port, destination port)* acts as a label of each particular instance.

Some well-known port numbers include:

- 21: File Transfer Protocol (FTP), used to exchange data between two network nodes.
- 22: Secure Shell (SSH), which permits exchange of encrypted data.
- 25: Simple Mail Transfer Protocol (SMTP), considered as the Internet standard for outgoing electronic mail.
- 80: HyperText Transfer Protocol (HTTP), used by the World Wide Web.
- 443: HTTPS, or HTTP Secure, which provides encrypted HTTP and secure identification of a network server through digital certificates.

In this chapter we define an *end-to-end (e2e) application-session*, or briefly session, as a cluster of bytes with the same source and destination network addresses, same transport protocol, and same destination port, such that the delay between two successive packets in the cluster is less than a predetermined threshold t. This is the simplest technique to guess applications; for example, sessions whose packets' destination port is 80 are regarded as web traffic. This definition captures the inner working of the process of listening. Due to its simplicity, we will use it here for doing statistical analysis of applications, but we acknowledge it has many drawbacks:

- Popular applications, such as peer-to-peer, streaming media and network gaming do not rely on *listening* through a predefined set of well-known ports. Rather, ports are assigned to applications when needed, and the transport protocol ensures that it does not assign the same port number to two processes, and that the numbers assigned are in the range of the private ports.
- There are applications with more than one port number assigned. For example, Adobe's Real Time Messaging Protocol (RTMP) for streaming audio, video, and data over the Internet. RTMP has three variations: the "vanilla" version which uses port number 1935, a second one which is encapsulated within HTTP to traverse firewalls through port number 80, and a third one which uses HTTPS and thus port number 443. Similarly, HTTP uses ports 80, 8000 and 8080.
- A single port number may be associated with traffic of more than one kind of application. A prominent example is port number 80, used widely by applications with the purpose to avoid firewalls.

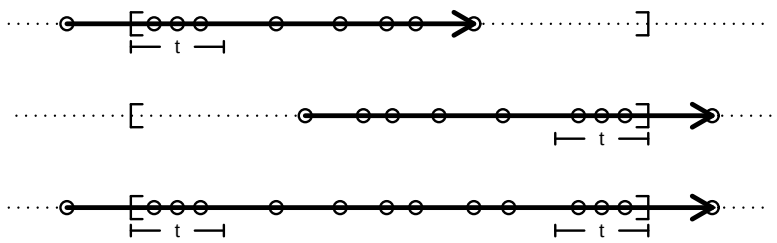

Figure 4.1: Three censoring types in data networks; here each packet is depicted as an oval, sessions are depicted as arrows, and the collection interval is the period within brackets. *Top:* Start-censoring. *Middle:* End-censoring. *Bottom:* Start/end-censoring.

- To make matters worse, even standard applications may run on non-standard ports in order to circumvent policy restrictions.

Keshav (1997) and Wetteroth (2001) provide a readable introduction to transport protocols and network ports.

4.2.2 Censored sessions

Censored observations in data networks can occur in three ways (see Figure 4.1)

Start-censoring: A session's first packet arrives before the start of the collection interval. Candidates are sessions whose first recorded packet lies within t time units of the start of the collection interval (see the definition of session in Section 4.2.1).

End-censoring: A session's last packet arrives after the end of the collection interval. Candidates are sessions whose last recorded packet lies within t time units of the end of the collection interval.

Start/end-censoring: A session satisfies both start and end censoring.

When a session presents any of these censoring types, its actual size and duration are only known to be above the observed value, i.e. they are right-censored. Furthermore, a censored session's rate is only known to be positive. Notice also that the number of censored sessions is unknown, as we can only identify censored candidates.

By increasing the measurement interval, we would reduce the number of start/end-censored sessions. However, this would not necessarily reduce the number of sessions that are only start- or only end-censored, and could also increase data handling and storage costs.

4.2.3 The data set

We present our analysis for an anonymized network trace captured at Equinix data center in San Jose, California, between 4:59 and 7:01 am, Coordinated Universal Time (UTC), of April 14, 2010. The data collection monitor is connected to a backbone link of a Tier 1 Internet Service Provider between San Jose and Los Angeles, California. As of June 2011, the data set is available upon request at the Cooperative Association for Internet Data Analysis (CAIDA)'s data repository at `http://www.caida.org/data/`. We have taken the part of the trace corresponding to TCP traffic going in a non-specified direction (dubbed as direction A). See the above CAIDA's website for more technical details about this

Table 4.1: Summary of network ports, ordered by numbers of bytes transmitted, with the associated application in parenthesis. 1: Processing and storage costs prevented us from using an hour of port 80; instead, we use the first $5min.$ of port 80 comprising a much smaller number of sessions. 2: Port 9050 sessions cannot be reconstructed with our definition; see the text for comments regarding this issue.

Port number (application)	Bytes/hr. transmitted	% of TCP bytes	Number of sessions
80 (HTTP)	968,891,084,629/hr.	85.20%	3,412,773/5 min.[1]
443 (HTTPS)	20,449,503,704/hr.	1.80%	3,296,257/hr.
9050 (TOR)	13,021,602,568/hr.	1.25%	[2]
25 (SMTP)	6,216,718,210/hr.	0.55%	1,215,546/hr.
1935 (RTMP)	6,112,271,508/hr.	0.54%	172,635/hr.

trace.

The raw data consists of 2,115,964,389 packet headers, from which we construct application sessions using a threshold between sessions of $t = 2s$ as in Chapter 3 and Sarvotham et al. (2005). This new data set has a large number of sessions belonging to several different network applications, a feature the Auckland's data set of Section 3.2.3 lacked.

We want to study ports with high usage in order to have large sample sizes. Table 4.1 shows the five most used ports by number of bytes, including their associated network applications. Observe that port 80 dominates the other four ports, so why should we bother analyzing those other ports? Although some ports, such as 25 and 443, represent a small percentage of the total TCP traffic, we still want to consider them because they are associated with indispensable

services like email and secure web, and there are millions of instances in which end-users use those services, as shown in the last column of Table 4.1. We also study port 1935 because it corresponds to streaming media and we believe this type of traffic may exhibit different features from web or email. We exclude port 9050 of our analysis because it consists of traffic generated by The Onion Router (TOR) network, a system that enables online anonymity by destroying identification of the ends of communication (see Dingledine and Mathewson, 2010), hence making our definition of session unworkable for this port.

Equinix's data set contains chronologically unordered packets which possibly are corrupted, making the computation of the sessions duration flawed. The ports we analyze here contain 3-8% of sessions with packets exhibiting this problem. Upon inquiry, CAIDA expressed surprise but offer no explanation for this anomaly. We left such sessions out of the analysis and Table 4.1's last column counts the remaining sessions.

4.3 Marginal distributions of S and D

We analyze the marginal distributions of S and D for each of the four ports of study. Historically, S and D have been modelled as having heavy tails and we look for evidence here. In Chapter 3 we applied Dietrich et al. (2002)'s statistic in the formal testing of $F \in \mathcal{D}(G_\gamma), \gamma > 0$; but that statistic does not account for censoring and we cannot use it here.

We first estimate the shape parameter γ_D of session durations. Although in principle we can use QQ and Hill plots based on the POT method, these do not consider censored sessions and thus may be biased toward lighter tails (smaller

γs). Hence, as an initial approximation, we use an estimator of γ_D by Beirlant and Guillou (2001) which assumes that start/end-censored sessions have the largest observed values. However, Beirlant and Guillou (2001)'s estimator assumes no other censoring type, and it also is not directly applicable to estimate the shape parameter γ_S of session sizes since start/end-censored sessions have the largest durations, but not necessarily the largest sizes. So we propose a maximum likelihood technique for estimating the shape parameter of a heavy-tailed distribution under censoring, and we use it to estimate γ_D and γ_S for all ports.

Why did not previous studies have problems with censoring? Graphically, start/end-censored session durations are the easiest to note, e.g. as vertical lines in QQ plots (see Figure 4.3 in Section 4.3.1) or also as big bars at the right end of histograms (not shown here). Old data sets, such as the one in Chapter 3, do not have start/end-censored sessions (see Table 4.2) because the applications that generate such large durations were previously scarce. Since we often use QQ plots for both detecting and estimating parameters of heavy tailed distributions, censoring was not frequently noticed. For instance, Table 4.2 shows port 1935 has the largest percentage of censored sessions after port 80, which is expected given that applications with long durations such as streaming media are associated with RTMP; but Auckland's data set from a decade ago only had a few hundred sessions from port 1935.

The percentage of censored sessions for port 80 is indeed higher than for port 1935, but we only harvested sessions from 5 minutes of port 80 traffic. This may account for the large percentage of censored sessions since it is conceivable that port 80 sessions started and ended outside such small measurement interval. Although we had the whole hour of packet headers, harvesting sessions from

port 80 was becoming difficult to process and store.

4.3.1 Beirlant-Guillou estimator for γ_D

Table 4.2: Number of censored candidates per network port (application) according to their type. The percentage of total number of sessions of a port these censored candidates represent appears in parenthesis. Censored candidates of Auckland's data set (Chapter 3) are also included.

Data set	Start/End-	Start-	End-
80 (HTTP)	$1,289(0.0377\%)$	$43,234(1.27\%)$	$34,357(1.01\%)$
443 (HTTPS)	$44(0.0013\%)$	$3,267(0.09\%)$	$2,057(0.06\%)$
25 (SMTP)	$6(0.0004\%)$	$1,144(0.09\%)$	$1179(0.09\%)$
1935 (RTMP)	$95(0.0550\%)$	$424(0.24\%)$	$517(0.30\%)$
Auckland's data set	$50(0.0000\%)$	$78(0.11\%)$	$0(0.17\%)$

Table 4.2 shows the number of sessions that might be censored, classified by the censoring type discussed in Section 4.2.2. For each port there is an unobserved random number $n_{S/E}$ of star/end-censored sessions whose recorded D is approximately $L_c = 3720s$ (the length of the collection interval) and greater than the duration of the remaining sessions. If n is the total number of sessions, those remaining $n - n_{S/E}$ sessions might include start- or end-censored sessions. How can we answer extreme value questions under censoring? Although censored sessions constitute a small percentage of the total number of sessions per port, but we wondered if this may cause our estimates to be biased toward lighter tails.

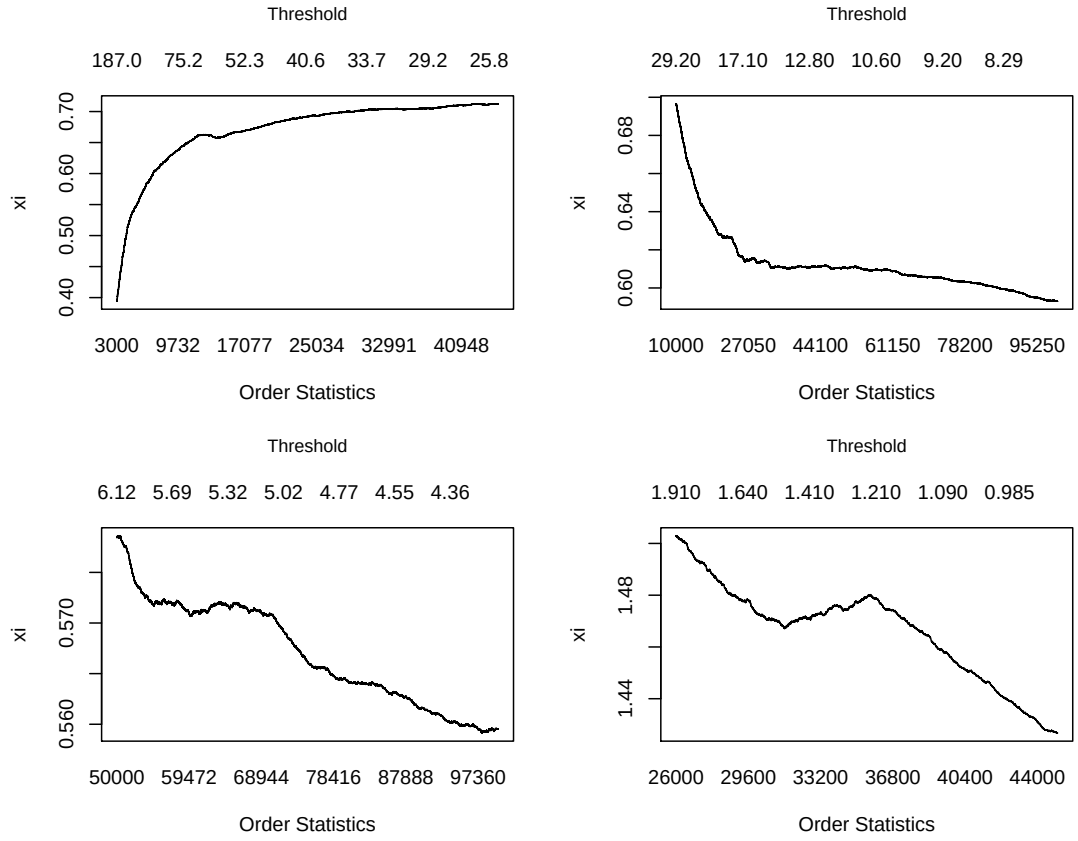

Figure 4.2: Hill plots of γ_D. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

Except for port 80, Hill plots of γ_D look very stable (Figure 4.2), but QQ plots (Figure 4.3) signal the presence of start/end-censored durations with values close to L_c. Table 4.3 shows that QQ estimates based on maximum likelihood do not coincide with Hill estimates. Which estimate is correct? Could they both be biased due to censoring?

As an initial approximation, assume that the start/end-censored candidates actually are censored and that their durations are exactly equal to L_c. For now, we ignore the start- and end-censored sessions.

This reduced problem can be formulated as follows: Let $Y_1, \ldots, Y_n$ be iid with common distribution $F \in \mathcal{D}(G_\gamma), \gamma > 0$. Assume that the first n_c upper order statistics are right-censored in such a way that for each of them we only know a "maximum observable value" M, but we do observe the actual values of the remaining $n - n_c$ statistics; therefore, our ordered sample is of the form

$$Y_{1:n} \leq \cdots \leq Y_{n-n_c:n} \leq Y_{n-n_c+1:n} = \cdots = Y_{n:n} = M. \tag{4.1}$$

Above a high threshold, a GPD QQ plot from such a sample would be linear up to the level M, at which it would be perfectly vertical at the final n_c points. Figure 4.3 shows this vertical effect in the GPD QQ plot of $\ln D$ for all the ports of study.

Since the Hill estimator can be viewed as a slope estimator in the GPD QQ plot above a threshold point, Beirlant and Guillou (2001) proposed an estimator of γ constructed by lumping the final n_c points heights at M, weighting these n_c pts appropriately, and adjusting weights used in the overall averaging of the censored and uncensored points. For $k > n_c$, the Beirlant-Guillou estimator is

$$\hat{\gamma}_{k,n} = \frac{1}{k - n_c} \left\{ \sum_{j=n_c+1}^{k} \ln \frac{Y_{n-j+1:n}}{Y_{n-k:n}} + n_c \ln \frac{M}{Y_{n-k:n}} \right\}, \tag{4.2}$$

which reduces to the Hill estimator in the absence of censoring ($n_c = 0$). Matthys et al. (2004) prove asymptotic normality of (4.2) as $k \to \infty, k/n \to 0$, $n \to \infty, n_c/k \xrightarrow{P} C \in [0, 1)$ and a second order condition; Beirlant and Guillou (2001) indicate via simulations that censoring should not exceed 5% for the estimator to work, which is the case for each port in our network data.

Figure 4.4 exhibits Beirlant-Guillou plots. Because censored sessions represent a very small percentage of the total number of sessions (Table 4.2), the second term in (4.2) barely contributes to the sum and Beirlant-Guillou are un-

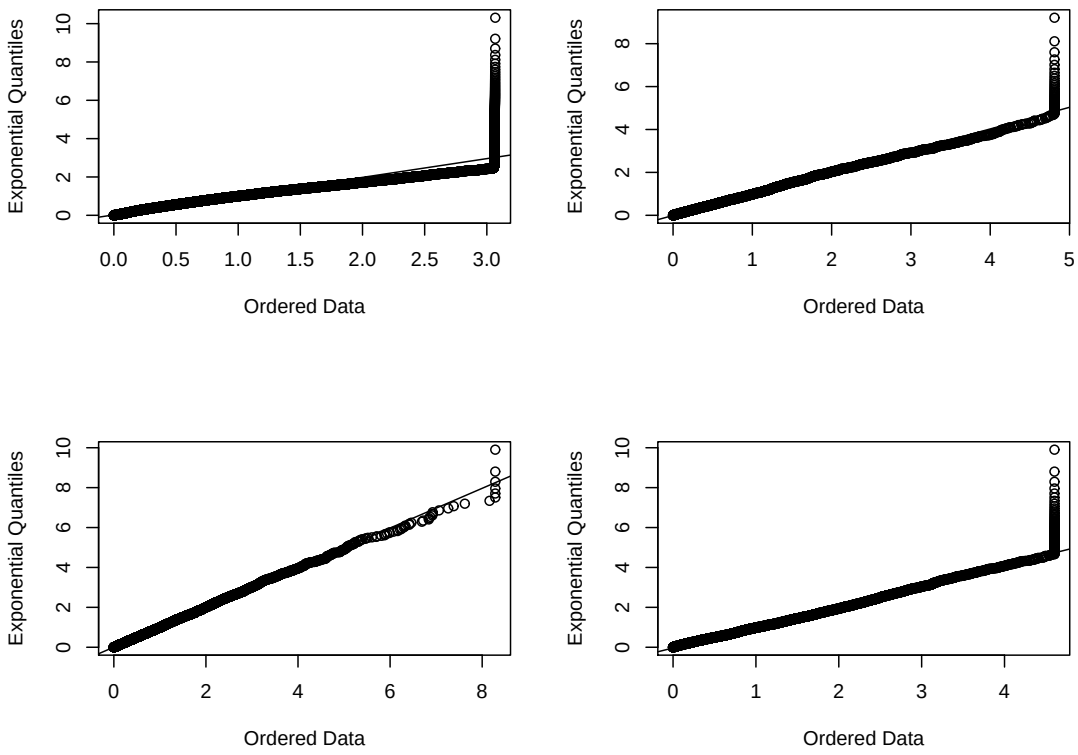

Figure 4.3: GPD QQ plots of log-durations, with $k \approx 10000$ upper order statistics used; start/end-censored sessions can be seen as a vertical line for all network ports. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

surprisingly similar to Hill plots. There are some ports showing wider regimes of stability than others. In particular, the plot for port 1935 (RTMP) does not exhibit much of a stability region and is hard to interpret but our best guess is $\gamma_D \approx 1.48$, indicating duration has an infinite first moment. According to the Beirlant-Guillou estimates, port 1935's durations have the heaviest tails, followed by ports 80 ($\gamma_D \approx 0.73$), 443 ($\gamma_D \approx 0.62$) and 25 ($\gamma_D \approx 0.57$). This result is expected given the corresponding application described in Section 4.2.1.

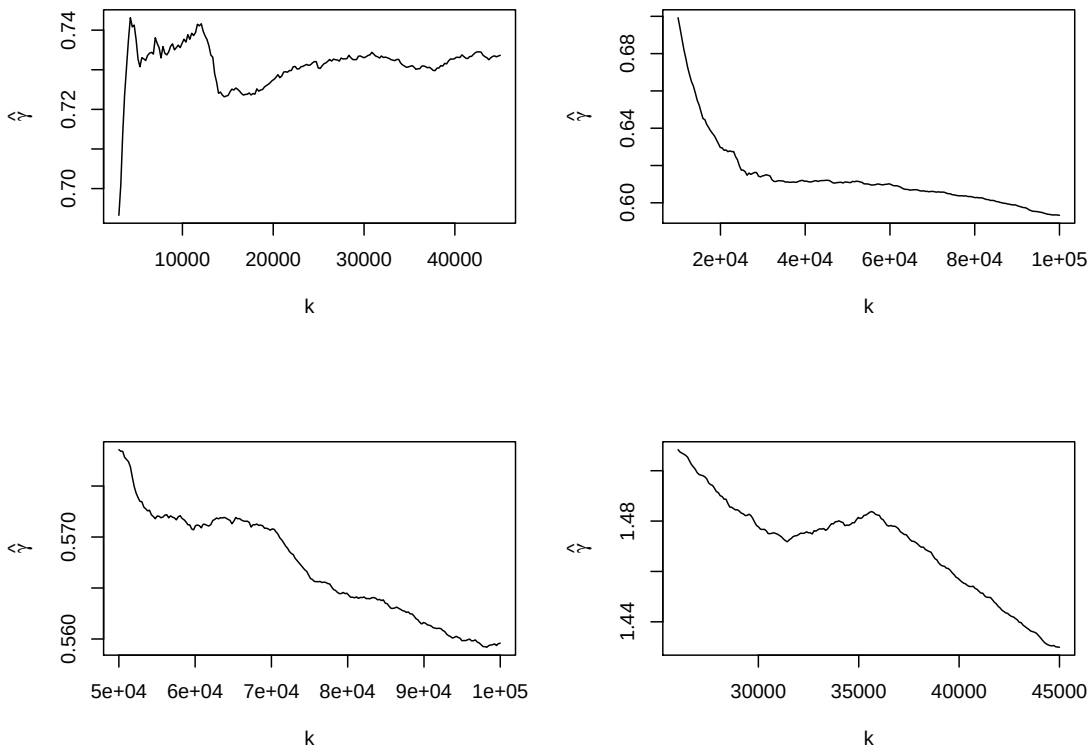

Figure 4.4: Beirlant-Guillou estimates γ_D, as a function of the number k of upper order statistics. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

We note one difference between Beirlant-Guillou and Hill plots: According to Table 4.2, port 80 (HTTP) has about 0.03% of star/end-censored sessions, which is small, but the Hill plot for γ_D looks less stable and puts the estimate of γ_D at a lower value than the Beirlant-Guillou plot, possibly indicating bias in the Hill estimator toward lighter tails. Table 4.3 in p. 110 summarizes these results.

Here we have ignored start- and end-censoring, which could also make Beirlant-Guillou estimator biased towards a lighter tail. We partially address

this issue in Section 4.3.2.

Can we use the Beirlant-Guillou estimator for γ_S? Unfortunately, Beirlant-Guillou plots are not stable for the shape parameter of S (not shown here). Part of the difficulty with sessions size is that S is not clearly constrained in the collection interval as D is, and the start/end-censored sessions do not necessarily have the largest S. As a result, censored sessions appear across all the list of order statistics of S and would contribute to the first term of the sum (4.2), even if we ignore the start- and end-censored sessions. Also, Hill plots are as bad as Beirlant-Guillou plots.

Then, what can we say about S? We can construct GPD QQ plots of log-sizes that exhibit linear trends (Figure 4.5 *Upper left* for port 80), or linearity within an interval right before the end of the tail (Figure 4.5). Relatively moderate changes in the number k of upper order statistics do not affect these linear trends. However, the QQ plots for the ports 443 (HTTPS), 25 (SMTP) and 1935 (RTMP) show that the tail of the maximum likelihood fit $GPD_{\hat{\gamma},\hat{\beta}}$ may be heavier than the tail of the distribution of S. One explanation is that censoring prevents us from observing the larger observations that would make up for this discrepancy. Given the corresponding application, it might also be that port 25's sizes do not have heavy tails, since outgoing email and attachments typically are bounded by the service provider. Next section also addresses the problem of estimating γ_S. These results summarized in Table 4.4.

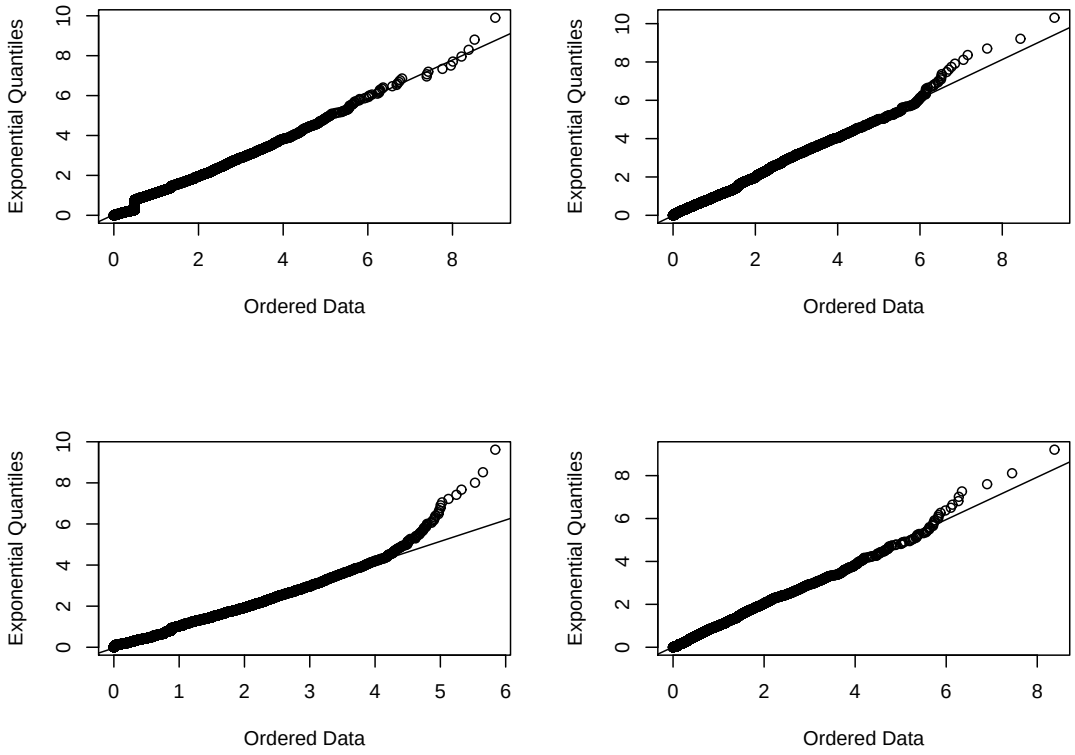

Figure 4.5: GPD QQ plots of log-sizes, with $k \approx 10000$ upper order statistics used. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

4.3.2 POT-MLE estimators for γ_S and γ_D in the presence of censoring

Let $Y_{1:n} \leq \cdots \leq Y_n$ be the order statistics of a sample from $F \in \mathcal{D}(G_\gamma)$. Suppose some data points are right-censored, i.e. they actually are larger than or equal to the observed value. Let C be the random set of indices of right-censored order statistics and $A = \{1, \ldots, n\}$. Asuming equality in (1.6), the log-likelihood under

censoring based on such a sample and a high threshold $\hat{u} = Y_{n-k:n}$ is given by

$$l(\gamma, \beta) = \sum_{j \leq k, j \in A \setminus C} \ln gpd_{\gamma,\beta}(Y_{n-j+1:n} - \hat{u}) + \sum_{j \leq k, j \in C} \ln \left(1 - GPD_{\gamma,\beta}(Y_{n-j+1:n} - \hat{u})\right),$$

and substituting the generalized Pareto distribution function $GPD_{\gamma,\beta}$ and its density $gpd_{\gamma,\beta}$ (see Section 1.2)

$$\begin{aligned} = & - |\{j \leq k\} \cup (A \setminus C)| \ln \beta - (1 + 1/\gamma) \sum_{j \leq k, j \in A \setminus C} \ln \left(1 + \gamma(Y_{n-j+1:n} - \hat{u})/\beta\right) \\ & - (1/\gamma) \sum_{j \leq k, j \in C} \ln \left[1 + \gamma(Y_{n-j+1:n} - \hat{u})/\beta\right], \end{aligned} \tag{4.3}$$

which can be maximized subject to the parameter constraints that $\beta > 0$ and $1 + \gamma(Y_{n-j+1:n} - \hat{u})/\beta > 0$ for $j \leq k$ (or equivalently $j = 1$). Numerically solving this optimization problem yields a maximum likelihood estimator of γ that depends on the number k of used upper order statistics, so in practice we make a plot of the estimate as a function of k and pick a value for which the graph looks stable.

Figure 4.6 shows the estimates of γ_D using this method. Observe that the estimate of γ_D for ports 1935 (RTMP), 443 (HTTPS) and 25 (SMTP) have been updated to $\gamma_D \approx 1.66, 0.7$, and 0.63 respectively, which suggests that start- and end-censoring effectively may have biased Beirlant-Guillou estimates toward lighter tails. Under this method, it also appears that port 1935's durations have the heaviest tails, followed by ports 443 and 25 in that order. Unfortunately, we cannot trust the estimate of γ_D for port 80 using this method, as Figure 4.6 exhibits its drastic change as a function of the number k of upper order statistics used.

Figure 4.7 repeats this analysis for the sessions size. Ports 80 (HTTP), 443 (HTTPS) and 25 (SMTP) exhibit narrow stable regimes at $\gamma_S \approx 1.48, 0.95$

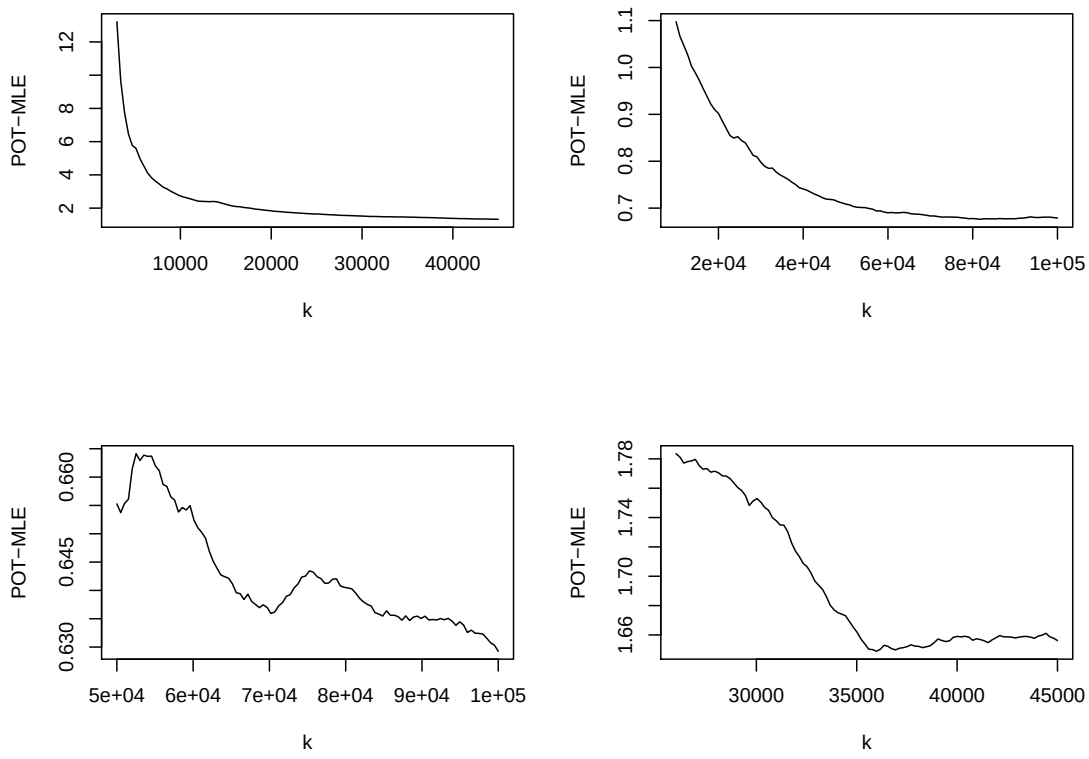

Figure 4.6: POT-MLE estimates of γ_D, as a function of the number k of upper order statistics. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

and 1.05, respectively. This method does not give a reliably estimate of γ_S for port 1935 (RTMP); it would be interesting to produce estimates for this latter port, as its associated application is the most different between the ones considered here. These pictures for the sessions size look unconvincing, and we think censoring may play a leading role in this problem.

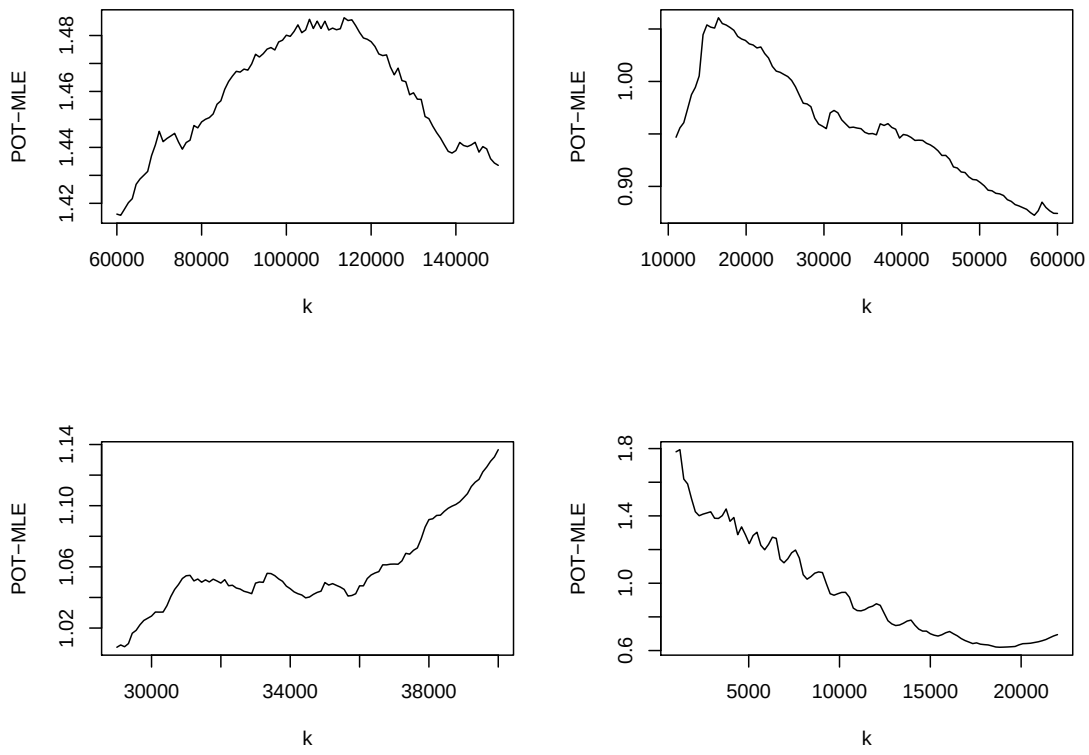

Figure 4.7: POT-MLE estimates of γ_S, as a function of the number k of upper order statistics. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

4.3.3 Summary of estimates of γ_S and γ_D

We summarize our estimates of γ_D and γ_S in Tables 4.3 and 4.4, respectively. Table entries are missing when the corresponding stability plot does not exhibit clear horizontal regimes.

Our QQ estimates are obtained by maximum likelihood using a number $k \approx 10000$ of upper order statistics for which the QQ plots look straight (see Figures 4.3 and 4.5). However, we mention here these estimates are stable only

Table 4.3: Summary of estimates of γ_D.

Estimator	QQ	Hill	Beirlant-Guillou	POT-MLE
Port 80 (HTTP)	0.35	0.69	0.73	Not reliable
Port 443 (HTTPS)	1.01	0.62	0.62	0.70
Port 25 (SMTP)	0.69	0.57	0.57	0.63
Port 1935 (RTMP)	1.06	1.48	1.48	1.66

Table 4.4: Summary of estimates of γ_S.

Estimator	QQ	POT-MLE
Port 80 (HTTP)	0.55	1.48
Port 443 (HTTPS)	0.92	0.95
Port 25 (SMTP)	0.91	1.05
Port 1935 (RTMP)	1.01	Not reliable

in a narrow range of k, and hence are less reliable than Hill, Beirlant-Guillou and POT-MLE estimates.

Also with the variable D, we see that the less censoring a method takes into account, the lower the estimate.

4.4 The Poisson property does not hold for network ports

Can we use the Poisson model as the generating mechanism for the individual network application sessions listed in Section 4.2.1?. To answer this question for

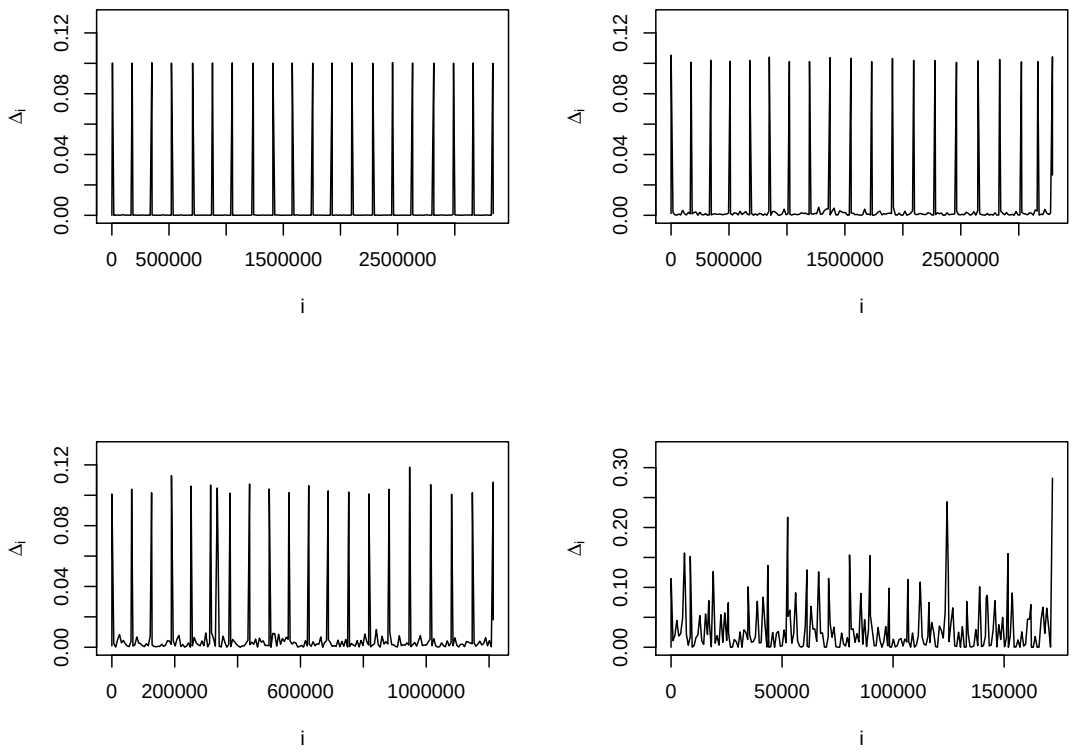

Figure 4.8: Time series $\{\Delta_i\}$. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

each port number separately, we study the interarrival times $\Delta_i = \Gamma_{i+1} - \Gamma_i$ of the sessions that are not censored; that is, those starting or ending away of the ends of the collection interval by $t = 2s$ (see Figure 4.1).

First, we investigate the stationarity of the series $\{\Delta_i\}$. For each of the four ports under study, the time series plots in Figure 4.8 show two distinct populations of interarrivals: small and large; the large interarrivals are shown as "spikes" and indicate sessions take longer than typical to arrive. Surprisingly, there are 3716 spikes for both ports 443 (HTTPS) and 25 (SMTP), each occurring

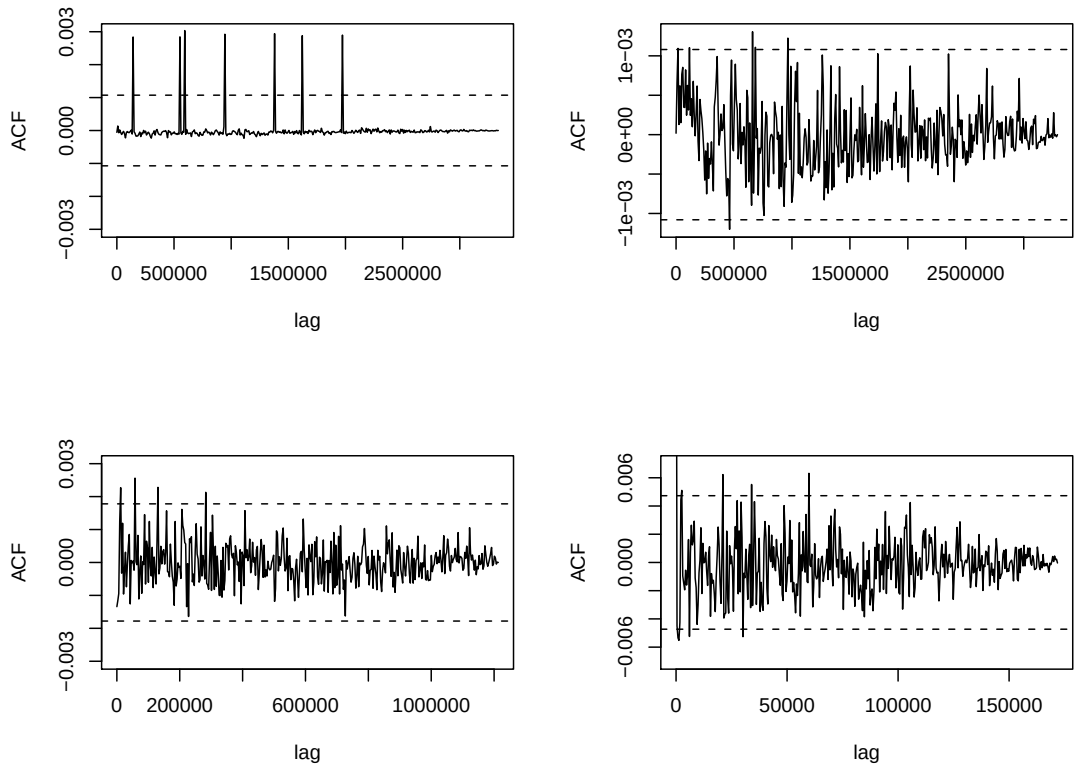

Figure 4.9: Sample autocorrelation functions of Δ_i. *Upper left:* Port 80 (HTTP). *Upper right:* Port 443 (HTTPS). *Lower left:* Port 25 (SMTP). *Lower right:* Port 1935 (RTMP).

within the same second across ports; for port 80 (HTTP), we only have 5 minutes of sessions so the number of large interarrivals is only 34, nevertheless they also occur within the same second as in the other ports. These large interarrivals may be associated with some sort of *network interruptions*; we will come back to this point later. For now, we note that the session interarrival times exhibit no obvious trend, even if we zoom into the parts of the plots excluding such network interruptions. So the series $\{\Delta_i\}$ looks stationary, and since it is bounded by the length of the collection interval, we can test for independence using the standard L_2 theory for confidence intervals (Brockwell and Davis, 1991).

Figure 4.9 shows sample autocorrelation plots of $\{\Delta_i\}$ for each of the four network ports. Confidence bounds for $\alpha = 0.05$ based on the standard Barlett's formula from classical time series analysis (Brockwell and Davis, 1991), are also shown. For all ports, we counted slightly less than 5% autocorrelations of the total lags outside the bounds, suggesting independence of $\{\Delta_i\}$ for these ports.

What can we say about the distribution of $\{\Delta_i\}$? As shown in Figure 4.8, for each port there are large interarrival times that may be associated with network interruptions. Further inspection shows that there are in fact two populations of interarrival times: those smaller than 0.05s, and the ones larger than 0.07s. This separation is so neat and it reinforces the idea that the large interarrivals may be caused by a network hardware mechanism.

The interarrivals occur in the following pattern: First a sequence of small $\{\Delta_i\}$ occur sequentially in chronological order until a single large Δ_i occurs; afterwards, another sequence of small $\{\Delta_i\}$ appears in chronological order, followed by a single large Δ_i, and the pattern repeats. There are 3716 large Δ_i for port 25 (SMTP), so the small Δ_i can be grouped into 3717 subpopulations such that all the small $\{\Delta_i\}$ of a given subpopulation occur in between two large Δ_i. Figure 4.10 *Left* contains a typical exponential QQ plot of one such population of small $\{\Delta_i\}$, with a striking straight line trend; this result replicates for the 3717 subpopulations of small $\{\Delta_i\}$. Given the aforementioned independence of interarrivals, we could describe the port 25 session arrival process as Poisson in between network interruptions.

We now compute the maximum likelihood estimate of the exponential parameter for each of the 3717 subpopulations of small $\{\Delta_i\}$, which also is an estimate of the arrival intensity of sessions in between network interruptions.

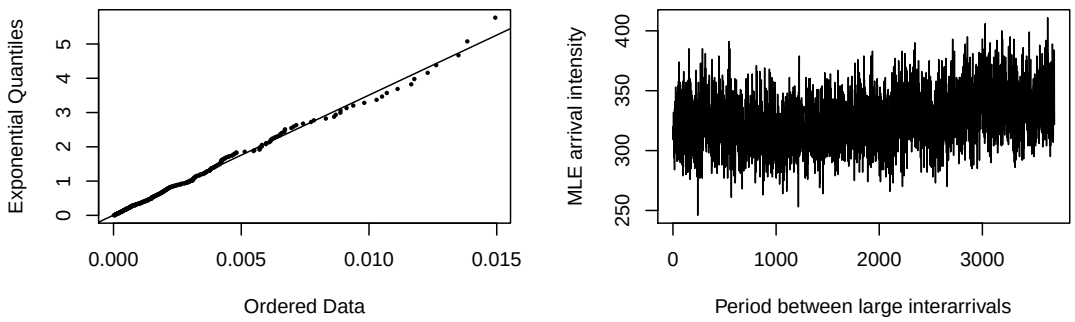

Figure 4.10: Distribution of port 25's $\{\Delta_i\}$. *Left:* Exponential QQ plot of a typical period between large interarrivals. *Right:* Plot of MLE arrival intensity as a function of the period between large interarrivals.

Figure 4.10 *Right* shows these MLE arrival intensities for all 3717 subpopulations in chronological order; we note that the plot looks jittery but shows no obvious trend, concentrating around 326 sessions/second. This suggests the session arrival process returns to a similar state after a network interruptions.

For ports 80 (HTTP) and 443 (HTTPS), we found independent but not exponential small $\{\Delta_i\}$. Port 1935 (RTMP) did not exhibit populations of small and large interarrivals as neatly separated as in the other ports. The more striking results for port 25 may be due to the fact that this port only carries SMTP (outgoing email) and is not corrupted by other applications like the other three ports.

4.5 Final remarks and conclusions

Although it is clear that censoring is present in data networks, it went unnoticed for Auckland's data set of Chapter 3. This is due partially to the age of the Auckland's data set: In 1999, only very few sessions lasted longer than an hour. After examining another CAIDA's dataset, we have found censoring has increased over the past year and requires more investigation.

Historic reasons made us search for heavy tails of S and D in this chapter, and we have shown compelling evidence that D has heavy tails for the ports 80 (HTTP), 443 (HTTPS), 25 (SMTP) and 1935 (RTMP). Table 4.3 collects estimates of γ_D using the various methods reviewed in this chapter.

We also were able to estimate γ_S for ports 80 , 443 and 25. For port 1935's size, we have unsuccessfully tried to fit other common distributions such as normal, lognormal and Gumbel.

For port 25, we found that the session arrival process could be simulated by Poisson processes between network interruption points. Similar catastrophe processes may be activating the sessions of port 80 and 443, but not of port 1935.

We also investigated the reliability of $(\ln S, \ln D)$ as classifiers of network ports, by implementing a simple $K-$means algorithm. Our initial analysis shows that $(\ln S, \ln D)$ are not good classifiers of ports, since only less than 10% of observations of each port are classified correctly. These findings suggest considering other types of application identification, or session variables in addition to (S, D).

BIBLIOGRAPHY

ARLITT, M. and WILLIAMSON, C. L. (1996). Web server workload characterization: The search for invariants (extended version). In *Proceedings of the ACM Sigmetrics international conference on Measurement and modeling of computer systems*. Philadelphia, PA, 126–137.

ATHREYA, K. B. (1987). Bootstrap of the mean in the infinite variance case. *The Annals of Statistics*, **15** 724–731.

BALKEMA, A. A. and DE HAAN, L. (1974). Residual life time at great age. *The Annals of Probability*, **2** 792–804.

BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics, John Wiley & Sons Ltd., Chichester.

BEIRLANT, J. and GUILLOU, A. (2001). Pareto index estimation under moderate right censoring. *Scandinavian Actuarial Journal* 111–125.

BILLINGSLEY, P. (1999). *Convergence of probability measures*. 2nd ed. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc.

BINGHAM, N. H., GOLDIE, C. M. and TEUGELS, J. L. (1987). *Regular variation*, vol. 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press.

BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*. 2nd ed. Springer-Verlag, New York.

CAO, J., CHEN, A., WIDJAJA, I. and ZHOU, N. (2008). Online identification of applications using statistical behavior analysis. *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008.* 1–6.

Cisco Systems, Inc. (2007). *Introduction to Cisco IOS NetFlow - A Technical Overview*. Cisco Systems, Inc., San Jose, CA. USA.

COLES, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics, Springer.

CROTTI, M., DUSI, M., GRINGOLI, F. and SALGARELLI, L. (2007). Traffic classification through simple statistical fingerprinting. *ACM SIGCOMM Computer Communication Review*, **37** 7–16.

CROVELLA, M. E. and BESTAVROS, A. (1997). Self-similarity in world wide web traffic: Evidence and possible causes. *IEEM/ACM Transactions on Networking*, **5** 845–846.

CSÖRGŐ, S., DEHEUVELS, P. and MASON, D. (1985). Kernel estimates of the tail index distribution. *The Annals of Statistics*, **13** 1050–1077.

CUNHA, C. R., BESTAVROS, A. and CROVELLA, M. E. (1995). Characteristics of www client-based traces. Technical report BU-CS-95-010, Computer Science Department, Boston University.

DAS, B. and RESNICK, S. I. (2011a). Conditioning on an extreme component: Model consistency and regular variation on cones. *Bernoulli*, **17** 226–252.

DAS, B. and RESNICK, S. I. (2011b). Detecting a conditional extreme value model. *Extremes*, **14** 29–61.

D'AURIA, B. and RESNICK, S. I. (2006). Data network models of burstiness. *Advances in Applied Probability*, **38** 373–404.

D'AURIA, B. and RESNICK, S. I. (2008). The influence of dependence on data network models. *Advances in Applied Probability*, **40** 60–94.

DAVIS, R. A. and RESNICK, S. I. (1984). Tail estimates motivated by extreme value theory. *The Annals of Statistics*, **12** 1467–1487.

DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, **52** 393–442.

DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction*. Springer-Verlag, New York.

DE HAAN, L. and PENG, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica. Journal of the Netherlands Society for Statistics and Operations Research*, **52** 60–70.

DE HAAN, L. and RESNICK, S. I. (1977). Limit theory for multivariate sample extremes. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **40** 317–337.

DE HAAN, L. and RESNICK, S. I. (1993). Estimating the limit distribution of multivariate extremes. *Stochastic Models*, **9** 275–309.

DE HAAN, L. and RESNICK, S. I. (1998). On asymptotic normality of the hill estimator. *Stochastic Models*, **14** 849–866.

DEHEUVELS, P., MASON, D. M. and SHORACK, G. R. (1993). Some results on the influence of extremes on the bootstrap. *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, **29** 83–103.

DEKKERS, A. L. and DE HAAN, L. (1989). On the estimation of the extreme-value index and large quantile estimation. *The Annals of Statistics*, **17** 1795–1832.

DHARMAPURIKAR, S., KRISHNAMURTHY, P., SPROULL, T. S. and LOCKWOOD, J. W. (2004). Deep packet inspection using parallel bloom filters. *IEEE Micro* 52–61.

DIETRICH, D., DE HAAN, L. and HÜSLER, J. (2002). Testing extreme value conditions. *Extremes*, **5** 71–85.

DINGLEDINE, R. and MATHEWSON, N. (2010). *The TOR Manual (https://www.torproject.org/docs/tor-manual.html.en)*. TOR project: Anonimity online, Walpole, MA.

DREES, H., DE HAAN, L. and LI, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions. *Journal of Statistical Planning and Inference*, **136** 3498–3538.

EMBRECHTS, P., KLUPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extreme Events for Insurance and Finance*. Springer-Verlag, Berlin.

GELUK, J., DE HAAN, L., RESNICK, S. and STĂRICĂ, C. (1997). Second-order regular variation, convolution and the central limit theorem. *Stochastic Processes and their Applications*, **69** 139–159.

GINÉ, E. and ZINN, J. (1989). Necessary conditions for the bootstrap of the mean. *The Annals of Statistics*, **17** 684–691.

GUERIN, C., NYBERG, H., PERRIN, O., RESNICK, S. I., ROOTZÉN, H. and STĂRICĂ, C. (2003). Empirical testing of the infinite source poisson data traffic model. *Stochastic Models*, **19** 151–200.

HALL, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society. Series B (Methodological)*, **44** 37–42.

HALL, P. (1990). Asymptotic properties of the bootstrap for heavy-tailed distributions. *The Annals of Probability*, **18** 1342–1360.

HEFFERNAN, J. E. and RESNICK, S. I. (2007). Limit laws for random vectors with an extreme component. *The Annals of Applied Probability*, **17** 537–571.

HEFFERNAN, J. E. and TAWN, J. A. (2004). A conditional approach for multivariate extreme values. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **66** 497–546.

HERNANDEZ-CAMPOS, F., NOBEL, A., SMITH, F. D. and JEFFAY, K. (2005). Understanding patterns of tcp connection usage with statistical clustering. In *IEEE MASCOTS*. 35–44.

HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, **3** 1163–1174.

HOHN, N., VEITCH, D. and ABRY, P. (2003). The impact of the flow arrival process in internet traffic. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP '03).*, vol. VI. 37–40.

HUANG, X. (1992). *Statistics of bivariate extremes*. Ph.D. thesis, Erasmus University Rotterdam, Postbus 1735, 3000DR, Rotterdam, The Netherlands.

HÜSLER, J. and LI, D. (2006). On testing extreme value conditions. *Extremes*, **9** 69–86.

JAIN, M. and DOVROLIS, C. (2005). End-to-end estimation of the available bandwidth variation range. In *Proceedings of the 2005 ACM SIGMETRICS International Conference on Measurement and Modeling of computer systems*. ACM, 265–276.

JIN, Y., BALI, S., DUNCAN, T. E. and FROST, V. S. (2007). Predicting properties of congestion events for a queueing system with fbm traffic. *IEEM/ACM Transactions on Networking*, **15** 1098–1108.

KAJ, I. and TAQQU, M. S. (2008). *In and Out of Equilibrium 2*, vol. 60 of *Progress in Probability*, chap. Convergence to Fractional Brownian Motion and to the Telecom Process: the Integral Representation Approach. Birkhäuser Basel, 383–427.

KALLENBERG, O. (1983). *Random measures*. 3rd ed. Akademie-Verlag, Berlin.

KARAGIANNIS, T., BROIDO, A., FALOUTSOS, M. and CLAFFY, K. (2004). Transport layer identification of p2p traffic. In *Proceedings of the 4th ACM SIGCOMM Workshop on Internet measurment*. 121–134.

KESHAV, S. (1997). *An Engineering Approach to Computer Networking; ATM Networks, the Internet, and the Telephone network*. Addison-Wesley, Reading, Mass.

KILPI, J. and NORROS, I. (2002). Testing the gaussian approximation of aggregate traffic. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. Session 2: modeling, ACM, Marseilles, France, 49–61.

KIM, M.-S., KANG, H.-J. and HONG, J. W.-K. (2003). *Towards Peer-to-Peer Traffic Analysis using Flows*, vol. 2867 of *Lecture Notes in Computer Science*, chap. 6. Springer, Berlin/Heidelberg, 55–67.

KIM, M.-S., WON, Y. J. and HONG, J. W.-K. (2005). Application-level traffic monitoring and an analysis on ip networks. *ETRI Journal*, **27** 22–42.

KINGMAN, J. F. C. (1993). *Poisson Processes*. Oxford Studies in Probability, Oxford University Press.

KORTEBI, A., MUSCARIELLO, L., OUESLATI, S. and ROBERTS, J. (2005). Evaluating the number of active flows in a scheduler realizing fair statistical bandwidth sharing. In *SIGMETRICS '05: Proceedings of the 2000 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*. ACM, 217–228.

KURTZ, T. G. (1996). Limit theorems for workload input models. In *Stochastic Networks: Theory and Applications* (F. Kelly, S. Zachary and I. Ziedins, eds.). No. 4 in Royal Statistical Society Lecture Note Series, Clarendon Press, Oxford, 119–139.

LEHMANN, E. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer Texts in Statistics, Springer.

LELAND, W. E., TAQQU, M. S., WILLINGER, W. and WILSON, D. V. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEM/ACM Transactions on Networking*, **2** 1–15.

LÓPEZ-OLIVEROS, L. and RESNICK, S. I. (2011). Extremal dependence analysis of network sessions. *Extremes*, **14** 1–28.

MAIOLINI, G., MOLINA, G., BAIOCCHI, A. and RIZZI, A. (2009). On the fly application flows identification by exploiting k-means based classifiers. *Journal of Information Assurance and Security*, **4** 142–150.

MASON, D. and TUROVA, T. (1994). Weak convergence of the hill estimator process. In *Extreme Value Theory and Applications* (J. Galambos, J. Lechner and E. Simiu, eds.). Kluwer Academic Publishers, Dordrecht, Holland, 419–432.

MATTHYS, G., DELAFOSSE, E., GUILLOU, A. and BEIRLANT, J. (2004). Esti-

mating catastrophic quantile levels for heavy-tailed distributions. *Insurance: Mathematics and Economics*, **34** 517–537.

MAULIK, K., RESNICK, S. I. and ROOTZÉN, H. (2002). Asymptotic independence and a network traffic model. *Journal of Applied Probability*, **39** 671–699.

MC NEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management*. Princeton Series in Finance, Princeton University Press, Princeton, NJ. Concepts, Techniques and Tools.

MIKOSCH, T., RESNICK, S. I., ROOTZÉN, H. and STEGEMAN, A. (2002). Is network traffic approximated by stable Lévy motion or fractional Brownian motion? *Annals of Applied Probability*, **12** 23–68.

MITRINOVIĆ, D. S. and VASIĆ, P. M. (1970). *Analytic Inequalities*, vol. 165 of *Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete*. Springer-Verlag, Berlin, New York.

MOORE, A. W. and ZUEV, D. (2005). Internet traffic classification using bayesian analysis techniques. In *SIGMETRICS '05: Proceedings of the 2005 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*. 50–60.

PARK, C., SHEN, H., MARRON, J. S., HERNANDEZ-CAMPOS, F. and VEITCH, D. (2006). Capturing the elusive poissonity in web traffic. In *Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '06), Sept. 11-14*. IEEE, 189–196.

PAXSON, V. and FLOYD, S. (1995). Wide area traffic: The failure of poisson modeling. *IEEM/ACM Transactions on Networking*, **3** 226–244.

PENG, L. (1998). *Second Order Condition and Extreme Value Theory*. Ph.D. thesis, Tinbergen Institute.

PICKANDS, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, **3** 119–131.

REISS, R.-D. and THOMAS, M. (2007). *Statistical Analysis of Extreme Values*. 3rd ed. Birkhäuser Verlag, Basel.

RESNICK, S. I. (1971). Tail equivalence and its applications. *Journal of Applied Probability*, **8** 136–156.

RESNICK, S. I. (1986). Point processes, regular variation and weak convergence. *Advances in Applied Probability*, **18** 66–138.

RESNICK, S. I. (1987). *Extremes Values, Regular Variation and Point Processes*. Springer-Verlag.

RESNICK, S. I. (2003). *SemStat: Seminaire Europeen de Statistique, Extreme Values in Finance, Telecommunications, and the Environment*, chap. Modeling Data Networks. Chapman-Hall, London, 287–372.

RESNICK, S. I. (2007). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer Series in Operations Research and Financial Engineering, Springer-Verlag, New York.

ROSIŃSKI, J. and RAJPUT, B. S. (1989). Spectral representations of infinitely divisible processes. *Probability Theory and Related Fields*, **82** 451–487.

SAMORODNITSKY, G. and TAQQU, M. S. (1994). *Stable non-Gaussian random processes: stochastic models with infinite variance*. Stochastic Modeling, Chapman & Hall.

SARVOTHAM, S., RIEDI, R. and BARANIUK, R. (2005). Network and user driven alpha-beta on-off source model for network traffic. *Computer Networks*, **48** 335–350.

SARVOTHAM, S., WANG, X., RIEDI, R. H. and BARANIUK, R. G. (2002). Additive and multiplicative mixture trees for network traffic modeling. In *ICASSP 2002: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. IV. IEEE, Signal processing society, Orlando, Florida, 4040–4043.

SHAKKOTTAI, S., BROWNLEE, N. and CLAFFY, K. (2005). A study of burstiness in tcp flows. In *Proceedings of the 6th International Workshop in Passive and Active Network Measurement, PAM 2005* (C. Dovrolis, ed.), vol. 3431 of *Lecture Notes in Computer Science*. Springer, Boston, MA, 13–26.

SHANE, A., DANIEL, L. and RICHARD, N. (2007). Extracting application objects from tcp packet traces. Tech. rep., University of Waikato, WAND Network Research Group.

TAQQU, M. S., WILLINGER, W. and SHERMAN, R. (1997). Proof of a fundamental result in self-similar traffic modeling. *ACM SIGCOMM Computer Communication Review*, **27** 5–23.

VAN DE MEENT, R. and MANDJES, M. (2005). Evaluation of 'user-oriented' and 'black-box' traffic models for link provisioning. In *Proceedings of the 1st Eu-*

roNGI Conference on Next Generation Internet Networks Traffic Engineering. IEEE, Rome, Italy, 380–387.

VAN DE MEENT, R., MANDJES, M. and PRAS, A. (2006). Gaussian traffic everywhere? In *ICC '06. IEEE International Conference on Communications, 2006.*, vol. 2. Istanbul, Turkey, 573–578.

WANG, Z. and LIU, J. (2007). Internet applications usage accounting based on flow. In *Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007.*, vol. 2. 113–116.

WETTEROTH, D. (2001). *OSI Reference Model for Telecommunications*. 1st ed. McGraw-Hill Professional Publishing.

WILLINGER, W. and PAXSON, V. (1998). Where mathematics meets the internet. *Notices of the American Mathematical Society*, **45** 961–970.

WILLINGER, W., PAXSON, V. and TAQQU, M. S. (1998). *A Practical Guide to Heavy Tails. Statistical Techniques and Applications*, chap. Self similarity and heavy tails: Structural modeling of network traffic. Birkhäuser Boston Inc., Boston, MA, 27–53.

WILLINGER, W., TAQQU, M. S., LELAND, W. E. and WILSON, D. V. (1995). Self-similarity in high-speed packet traffic: Analysis and modeling of ethernet traffic measurements. *Statistical Science*, **10** 67–85.

WILLINGER, W., TAQQU, M. S., SHERMAN, R. and WILSON, D. V. (1997). Self-similarity through high variability: Statistical analysis of ethernet lan traffic at the source level. *IEEM/ACM Transactions on Networking*, **5** 71–86.

ZHANG, Y., BRESLAU, L., PAXSON, V. and SHENKER, S. (2002). On the characteristics and origins of internet flow rates. ACM Sigcom 2002 Conference, Pittsburgh, Pa; August 19-23.