TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY
Julkaisu 823 • Publication 823

Mikko Laurikkala

Goodness-of-Fit Tests and Heavy-Tailed Distributions in Network Traffic Data Analysis

Tampere 2009

Tampereen teknillinen yliopisto. Julkaisu 823
Tampere University of Technology. Publication 823

Mikko Laurikkala

Goodness-of-Fit Tests and Heavy-Tailed Distributions in Network Traffic Data Analysis

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Sähkötalo Building, Auditorium S2, at Tampere University of Technology, on the 21st of August 2009, at 12 noon.

Tampereen teknillinen yliopisto - Tampere University of Technology
Tampere 2009

ISBN 978-952-15-2191-1 (printed)
ISBN 978-952-15-2233-8 (PDF)
ISSN 1459-2045

Abstract

Network management system is a vital part of a modern telecommunication network. The duties of the system include, among other things, fault management, configuration management, and performance management. For these purposes the network management system collects vast amounts of data, the processing and analysis of which has developed into a whole discipline. Network traffic data analysis involves, for example, change detection, prediction, and modelling.

This thesis concentrates on network traffic data analysis with statistical tools, goodness-of-fit tests in particular. Instead of artificially generated data, data sets collected from real networks serve as case examples. Since real network data fit poorly to analytical distributions or textbook examples, Monte Carlo simulation is used for modelling the properties of the data.

The various quantities measured from telecommunication networks reportedly exhibit heavy-tailed distributions. Heavy-tailed distributions possess special features (such as infinite variance) that make them problematic for statistical analysis as well as network management. This is why heavy-tailed distributions are one of the premises of this work.

The network management system usually does not allow tailoring the measurements for a specific purpose but the analysis has to adapt to the data available. A histogram is one of the most popular means to compress data, that is, the data from the system often come as a histogram. This work develops a method for change detection of histogram data.

Furthermore, classical goodness-of-fit tests are largely inadequate for network traffic data. In addition to heavy-tailed distributions, the huge amount of data causes problems. This thesis collects several test statistics proposed in the literature for testing heavy-tailed distributions. Their usefulness is assessed through a power study, where a scenario of true traffic change detection is created. According to the results, the plain median outperforms all the more complicated test statistics in change detection. A suitable sample size is sought with a similar

power study, because the large amount of data may easily ruin the feasibility of the test.

Some sources cite predictability as an advantage of heavy-tailed distributions, but this feature has never been exploited. This thesis first generalizes the predictability to the time-continuous domain and then develops it further to a model that tries to predict traffic volume. However, the usefulness of the predictability remains limited, because several assumptions have to be made that do not necessarily hold in real network applications.

Preface

Some friends have asked me when I started preparing my doctoral thesis. I'm not sure, but now it is ready. During the years, I have participated in several research projects as well as teaching and finally found the right track to finish my thesis. I am grateful for the funding that I have received from research projects, TUT and PMGS.

I wish to thank my supervisor, Prof. Hannu Koivisto, for guidance and assistance in the research. I also thank Prof. Riku Jäntti and Dr. Sampsa Laine for many valuable comments and co-operation in the pre-examination phase.

I carried out most of the research in the Institute of Automation and Control. At the final stage however, we merged with our neighbours to constitute the Department of Automation Science and Engineering, and I got plenty of new, excellent colleagues. I am grateful to all of you for the pleasant environment, though I cannot mention all the names here.

M.Sc. Jari Seppälä has helped me in many things throughout the years. We even laid bets on the completion of this thesis; unfortunately neither of us remembers who wins now that it is ready. Senior researchers Dr. Matti Vilkko and Dr. Terho Jussila have given me useful advice. M.Sc. Johannes Hannila, M.Sc. Risto Silvola, and M.Sc. Timo Lehto helped me in collecting measurement data and invading into the complex world of telecommunications. Many helpful thoughts have emerged also in discussions with M.Sc. Pietari Pulkkinen, M.Sc. Aino Ropponen, M.Sc. Mariaana Savia, M.Sc. Pekka Kumpulainen, M.Sc. Heimo Ihalainen, and Prof. Risto Ritala.

The warmest thanks are due to my wife Heli and my daughters Nella and Milla. At their pre-school age, the girls never understood what dad is doing at work. This thesis hardly makes it any clearer, but now they at least know what a dissertation is.

Tampere, June 2009

Mikko Laurikkala

Contents

1 Introduction **1**

1.1 Short history of traffic self-similarity and heavy-tailedness 2

1.2 Role of statistical methods 5

1.3 Network management 7

1.3.1 Network management protocols 9

1.4 Contributions 12

1.5 Structure 13

2 Mathematical methods **15**

2.1 Random variables 16

2.2 Density and distribution functions 16

2.2.1 Distribution estimation 18

2.2.2 Exponential distribution 20

2.2.3 Heavy-tailed distributions 20

2.2.4 Self-similarity 21

2.2.5 Long-range dependence 23

2.3 Statistical testing 24

2.3.1 Discrete distributions and statistical testing 27

2.3.2 Monte Carlo simulation 29

2.3.3 Goodness-of-fit tests 30

2.3.4 Power 33

3 Goodness-of-fit tests and network traffic data **35**

3.1 Normal or lognormal distribution 36

3.1.1 Visual analysis 37

3.1.2 Anderson-Darling test for normal and lognormal distributions 38

3.2 Pareto distribution 40
3.2.1 Visual analysis 40
3.2.2 Anderson-Darling test for censored Pareto distribution . . 43
3.2.3 Sample size considerations 45
3.3 Conclusion 47

4 Change detection of discrete data **49**
4.1 Comparing samples of discrete distributions 51
4.1.1 Comparing port number histograms 51
4.2 Change detection of histogram data 57
4.2.1 Case example: interarrival times of GPRS packets 59
4.3 Conclusion 63

5 Test statistic study **67**
5.1 Introduction of test statistics 68
5.2 Selection of data 70
5.3 Test arrangement 72
5.4 Results 73
5.4.1 HTTP as null hypothesis 73
5.4.2 Gnutella as null hypothesis 74
5.4.3 Change detection between Gnutella and Kazaa 76
5.5 Conclusion 77

6 Sample size study **81**
6.1 Example: information content of a large sample 83
6.2 Test arrangement 83
6.3 Results 85
6.4 Conclusion 88
6.4.1 Guidelines for change detection with goodness-of-fit tests 89

7 Heavy-tailed distributions in traffic prediction **91**
7.1 Heavy-tailed distributions 94
7.2 Predicting flow durations 95
7.2.1 Distribution of flow age 97
7.2.2 Predictability of the renewal process 99
7.3 Applicability of the prediction 104
7.3.1 Renewal process 105
7.3.2 Alternating renewal process 107

7.3.3 No renewals 108
7.4 Conclusion 109

8 Conclusion **111**
8.1 Results 111
8.2 Discussion 113

List of symbols

$=_d$	is distributed equally
1_a	1 if a is true, 0 otherwise
a	tail index, shape parameter of Pareto distribution
A^2	Anderson-Darling statistic
$A^{2'}$	modified Anderson-Darling statistic
b	scale parameter of Pareto distribution, bin center
c	number of possible values of a discrete random variable, number of categories in a histogram
d	effect size
D	flow lifetime
E	expected frequency
$\mathrm{E}\{X\}$	expected value of random variable X
$f(x)$	probability density function
$f_n(x)$	histogram
$F(x)$	cumulative distribution function
$F_X(x)$	cumulative distribution function of random variable X
$F_{\mathrm{Par}}(x)$	cumulative distribution function of a Pareto-distributed random variable
$F_{\exp}(x)$	cumulative distribution function of an exponentially distributed random variable
$F_n(x)$	empirical cumulative distribution function of n observations
$\overline{F}(x)$	complementary cumulative distribution function
h	histogram bin width, prediction horizon
H	Hurst parameter

H_0	null hypothesis
H_1	alternative hypothesis
k	constant in a heavy-tailed distribution, delay of autocorrelation function
k_0	number of simulated null distribution samples
k_1	number of simulated test samples
K, U, V, W, Z, A, B, C	test statistics
KL	Kullback-Leibler distance
l	censoring threshold
m	number of observations in a category, sample mean
n	sample size
N_0	number of idle terminals
N_1	number of active flows
O	observed frequency
p_i, q_i	probabilities of a discrete distribution
P	P-value of a test statistic
$P(\theta)$	power of a test
P_{11}	probability that a flow remains open
$\Pr\{\cdot\}$	probability
s^2	sample variance
t	time
T	generic test statistic, flow age
T^*	value of test statistic T
U	uniform random variable
X, Y, Z	random variables
$\mathbf{X}$	data set
$X_{(i)}$	ith order statistic of X
$\overline{X}$	sample mean
$\overline{X}_i$	mean of the ith quartile
α	significance
$\beta(\theta)$	probability of type II error
θ	generic distribution parameter, location parameter of exponential distribution

μ	mean, scale parameter of exponential distribution
σ	standard deviation
σ^2	variance

List of abbreviations

A-D	Anderson-Darling
AR	Autoregressive
ARMA	Autoregressive Moving Average
CAC	Call Admission Control
ccdf	complementary cumulative distribution function
cdf	cumulative distribution function
CMIP	Common Management Information Protocol
DNS	Domain Name System
ecdf	empirical cumulative distribution function
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HTTP	Hypertext Transfer Protocol
IANA	Internet Assigned Numbers Authority
ICMP	Internet Control Message Protocol
IP	Internet Protocol
IRC	Internet Relay Chat
ISO	International Organization for Standardization
ITU-T	International Telecommunication Union, Telecommunication Standardization Sector
KL	Kullback-Leibler
K-S	Kolmogorov-Smirnov
LAN	Local Area Network
LRD	Long-range Dependence
MC	Monte Carlo (simulation)
MIB	Management Information Base

MIT	Management Information Tree
OAMP	Operations, Administration, Maintenance and Provisioning
OSI	Open Systems Interconnection
p2p	peer-to-peer
pdf	probability density function
PIT	Probability Integral Transform
QoS	Quality of Service
RLS	Recursive Least Squares
RMON	Remote Network Monitoring
SMI	Structure of Management Information
SNMP	Simple Network Management Protocol
TCP	Transmission Control Protocol
TMN	Telecommunications Management Network
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
WWW	World Wide Web

Chapter 1

Introduction

A distinctive feature of today's Internet is an explosive growth. At the beginning of the new millennium, the number of hosts with a registered IP address was 72 million. In June 2006, the number had grown 6-fold to 439 million. The respective numbers of WWW servers were even more drastic: 10 million in 2000, 88 million in 2006. [121]

In this thesis however, the emphasis is not on the hosts but the data transmitted over the Internet and other networks. It would be interesting to present statistics of the growth in data volumes, but estimating the amount of data in every nook and cranny of computer networks is hardly possible. All Internet exchange points collect statistics on volumes passing through their ports, but only a fraction of all traffic goes up to these exchange points.

Tanenbaum [109] defines a computer network as "a collection of autonomous computers interconnected by a single technology". According to this definition, company intranets, mobile phone networks, and even home PCs connected to each other are computer networks. Kurose and Ross [62] cite broadband residential access, mobile Internet, and peer-to-peer applications as killer applications that influence the nature of computer usage. Each of them not only brings a substantial increase into the traffic volume but also adds its own flavour to the traffic. These recent features of computer networks fortify the starting point for this study: analysis of traffic transmitted by computer networks is a basis for developing the

networks.

The objective of this thesis is to explore the possibilities of statistical methods in network traffic data analysis. The wide diversity of statistical methods is utilized quite narrowly in the thesis; hypothesis testing and goodness-of-fit testing receive particular attention. The nature of telecommunication network traffic coerces the emphasis to heavy-tailed distributions and other characteristics of the traffic data. Another focus is change detection of the traffic.

This introductory chapter first reviews some related fields, mainly heavy tails and network management. Then, the contributions and the structure of the thesis are presented.

1.1 Short history of traffic self-similarity and heavy-tailedness

Intuitively, self-similarity appears as similar structures on different scales. The pioneer of the field, Benoit B. Mandelbrot mentions *statistical* self-similarity as some kind of a special case of self-similarity [72]. Park and Willinger [84] divide the phenomenon into *deterministic* and *stochastic* self-similarity. While deterministic self-similarity has some impressive applications like fractal geometry, statistical or stochastic self-similarity has proven useful in modelling various processes in, for example, hydrology, finance, and physics [2]. This thesis concentrates on statistical self-similarity in network traffic data, which the term self-similarity henceforth refers to without further specifications. Although self-similarity can be applied also to spatial data [15], the focus is on time domain.

Whereas fractal curves repeat the same shape in different sizes, a statistically self-similar process repeats distributional similarities on different scales. In a time series this appears as burstiness that does not fade out by averaging even on a long time scale. This kind of burstiness is particularly typical to Internet traffic. [26]

Phenomena related to self-similarity include heavy-tailed distributions and long-range dependence (LRD). Heavy-tailed distributions are known to cause self-

similarity [117], thus they can be used to model self-similarity. The name comes from the slowly decaying tail of the cumulative distribution function

$$\Pr\{X > x\} \sim x^{-a}, \tag{1.1}$$

as opposed to the exponentially decaying tails

$$\Pr\{X > x\} \sim e^{-x} \tag{1.2}$$

of normal and exponential distributions. The heavy tail incurs non-negligible probabilities of extremely large values. These large values then cause bursts, overflows, and other trouble to network management.

Long-range dependence causes present values to depend on values extremely far in the past. In telecommunication networks, these phenomena appear as, for example, long periods of traffic activity above the mean or bursts of different durations. Consequently, self-similarity in network traffic has attracted a lot of research during the last decades.

As early as 1981, Pawlita [85] collected and analyzed data from some telecommunication systems. His findings of bursty traffic and diverse user applications were amazingly similar to the nature of today's network traffic. He also outlined requirements for future measurement systems.

Leland *et al.* collected their famous Bellcore Ethernet traces in the early 1990's; they first found the traces to contain bursts over many time scales [68] and then declared the phenomenon to be statistical self-similarity [66, 67]. A common conclusion of these and many other studies [87, 16] was that the traditional methods used for traffic modelling turned out to be insufficient.

Efforts to explain self-similarity in network traffic followed. The first analyses used Ethernet, FTP, and video traffic data [42], but the great growth of the World

Wide Web was just around the corner. As soon as 1996, Crovella and Bestavros published results on self-similarity in the WWW [26]. They found the document sizes in WWW to follow a heavy-tailed distribution, causing self-similarity in traffic volumes. Another suspect for arousing self-similarity was the control mechanism of Transmission Control Protocol (TCP) [38].

Several branches of research have diverged after the pioneering work. One of them is a practical approach, where metrics and characteristics are sought that describe the degree of heavy tails or self-similarity. Graphical methods [15, 28], Hill estimator [49] and Whittle estimator [15] have been used to calculate the tail index of a distribution or the Hurst parameter, a measure of self-similarity. A comparison of some estimators is provided in [110].

Another focal topic among network traffic research is modelling and analysis. In addition to the review in [4], wavelets deserve a mention. Abry and Veitch [3] developed a wavelet-based method for analysis of long-range dependence and used it for estimating the Hurst parameter as well as the fractal nature of traffic. They also implemented a real-time version of the the method [95]. Huang *et al.* [52] used wavelets for detecting network performance problems.

Prediction is one of the main targets in all traffic modelling, also in Chapter 7 of this thesis. Norros [80] used fractional Brownian motion to model the self-similarity and to create short-time predictions of the traffic. His model contained two parameters in addition to the Hurst parameter, namely mean input rate and variance coefficient.

The peculiar distributions of network traffic have lead to another way of classifying it. The mice and elephants terminology [75] is a popular metaphor. Mice are tiny flows carrying only a few packets of data, such as web page requests or routing information. While mice constitute a vast majority of the flows, rare elephant flows transfer a significant portion of the payload. Mice and elephants sometimes refer to byte counts, sometimes to flow durations. Brownlee and Claffy [22] also introduced dragonflies and tortoises to describe flow durations.

Since the first seminal papers, heavy-tailedness and self-similarity have been found in a variety of networks. The dawn of the mobile Internet does not seem to dissolve the phenomenon: self-similarity has been reported in WAP [71] as well

as UMTS [55] traffic. Peer-to-peer applications, another recent boom of the Internet, not only account for a notable portion of the traffic [63] but also exhibit self-similarity [57]. The consensus on self-similar network traffic has been quite strong in spite of some contradictory opinions soon after the first results [36]. Lately, Karagiannis *et al.* [58] questioned the omnipotence of self-similar models by asking, "Why should traffic be an exception to the Internet's diversity?" Still, even geometric self-similarity in the topology of the Internet has been reported [34].

The three terms — heavy-tailed distributions, self-similarity, and long-range dependence — are closely related but not equivalent. This thesis is engrossed in heavy-tailed distributions even up to its title. The two other concepts are entailed in a minor role whenever needed.

1.2 Role of statistical methods

Paxson [86] blamed network traffic studies, including his own, for inadequate use of statistics. Ten years later, Hajji [45] expressed similar opinions by considering the effort to analyze data too small compared to the work of collecting the data. Inevitably, there is a growing need for statistically competent analysis methods for network traffic.

A discipline called *network traffic data analysis* acts as a connector between telecommunications and data analysis (Figure 1.1). Telecommunications is clearly the application field, while data analysis as a science provides tools for mining the huge amounts of data collected. This thesis takes its place in the field of network traffic data analysis or, to be more exact, in a subset where statistical methods and network management intersect.

The American Heritage Dictionary [6] defines *statistics* as "the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling." The first half of the definition involves almost any data analysis method, but the complement about population and sampling limits the term statistics to what is usually taught

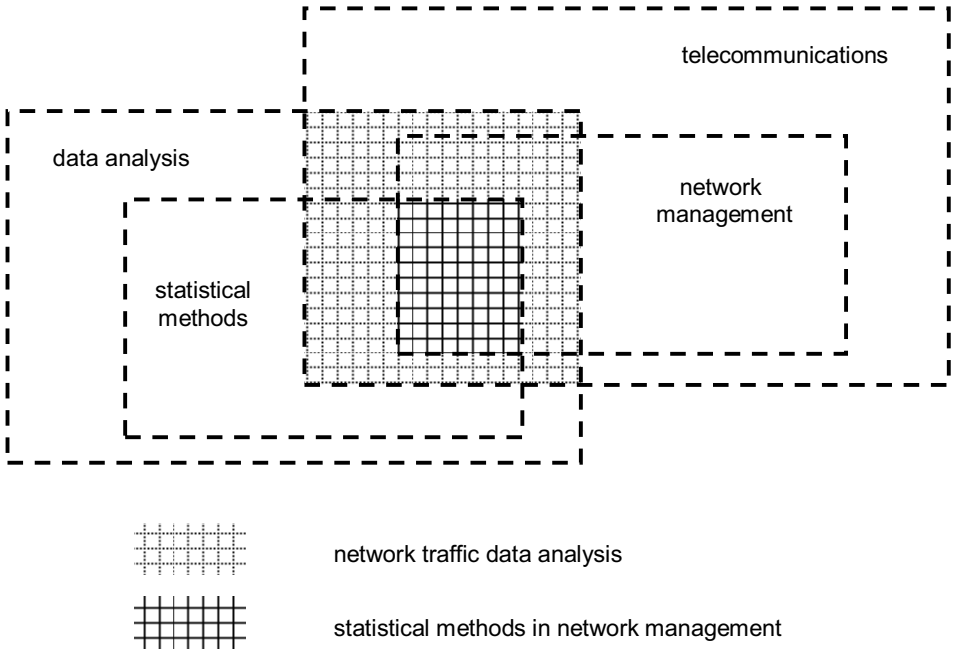

Figure 1.1: The field of the thesis in the context of telecommunications and data analysis.

in classes. Thus, statistical methods are considered here only a small part of the data analysis repertoire.

Among statistical methods, goodness-of-fit tests receive particular attention in this thesis. They do not belong to the most popular methods of network traffic data analysis, see Chapter 3 for a brief review. The thesis uses goodness-of-fit tests in several case examples and explores new ways to utilize them, especially in change detection.

A goodness-of-fit test examines how well a sample of data fits a given distribution [31]. In the case of change detection, two samples can be compared with each other or to a reference distribution. Compared to other possible change detection methods, goodness-of-fit tests have a strong theoretical background from statistics. They have been studied for decades, and the choices made by the data analyst (significance level, for example) are mathematically well understood. Many

heuristic methods can be used for change detection, but such a method inevitably needs plenty of tailoring to the case. For example, a simple threshold easily detects changes in traffic volumes, but how to set the threshold? A neural network can be trained to separate anomalies from the data, but designing and training the network is not straightforward. A statistical test may need as much tuning as the threshold-based method, but it uses proven methodology and tells the result in a well-known manner: the hypothesis can or cannot be rejected with a certain significance.

1.3 Network management

When a telecommunication network is working well, a user hardly notices the network. Protocols, requests, and file transfers are hidden under the application layer, and everything occurs almost without any delays. However, the network is not running on its own; its correct action requires constant attention and care. Network management is a broad term for this attention, even though an unambiguous definition of network management does not exist. A small LAN as well as the global Internet needs network management. This section gives a coarse overview of the elements of network management.

Effective network management optimizes the operational capabilities of the network. Goals of the management include keeping the network operating at its peak performance, informing the operator of threatening faults and their causes, maintaining network security, and collecting data on network usage [40].

Network management can be defined as operations, administration, maintenance, and provisioning (OAMP) of network and services. Administration means attending to the high-level goals, policies and procedures of network management. Network maintenance takes care of both installation and repairs of facilities and equipment. By means of provisioning, the network is planned and delivered according to the customers' needs. [108]

One of the targets of network management is to ensure that the users of a network receive the services with the Quality of Service (QoS) that they expect [108]. Un-

til recently, the Internet has offered equal service, referred to as best effort service, to all users. The term Quality of Service has meant properties like availability, reliability and integrity; or in more measurable terms, bandwidth, delay and jitter (variance of delay). The growing demand of business and quality-critical applications has turned QoS into class-based thinking: Users willing to pay more for the quality get wider bandwidth, shorter delays and less jitter. The traditional best effort service is left to the users who do not require premium class quality and thus do not want to pay for it.

The role of network traffic data analysis in QoS is to master the stochastic properties inherent to network traffic. As the quality parameter — for example, delay or available bandwidth — always is a random variable, it varies inside the acceptable region and, unfortunately, also outside it. By means of data analysis, we can make deductions about changes, true values, and causes and effects of quality parameters.

ITU-T lists five management functional areas of network management [90]. These five areas are commonly referred to in the literature, often related to the OAMP functions. In the following list, the functional areas are introduced from the standpoint of data analysis in general and this thesis in particular.

Fault management logs, detects and responds to fault conditions in the network [62]. Specifically, fault detection is an obvious application of data analysis. Sudden, abrupt faults are easily detected, but distinguishing a slow drift towards a possible fault from the normal operation calls for advanced analysis. This thesis covers both change detection and prediction.

Performance management contains quantifying, measuring, analyzing and controlling the performance of the network [62]. The performance may be quantified in a service level agreement (SLA) using statistical terms, for example: “On average, bandwidth availability will not be less than 30 %.” [70] Detecting whether the defined condition holds is a problem of hypothesis testing. Goodness-of-fit tests, one of the focuses throughout this thesis, can be used for analyzing changes in the performance.

Security management involves intrusion detection, an intensely studied topic. Intrusion and other anomaly detection can use statistical methods, including goodness-of-fit tests suitable for change detection.

Configuration and accounting management are further areas of network management. They are not related to the methods presented in this thesis as directly as the other three areas. Yet, data analysis may well be needed in, for example, accounting management.

1.3.1 Network management protocols

Several network management protocols have emerged from different communities. The oldest, most popular and simplest is the Simple Network Management Protocol, SNMP. In addition to the protocol itself, SNMP contains the Management Information Base (MIB) and Structure of Management Information (SMI).

In SNMP-based network management, an agent acts in a network element and maintains management information in MIB, a local database. SMI defines the structure of MIB and allowed data types. The network has one or more management servers collecting information from the agents with requests and traps following the SNMP protocol (Figure 1.2). It is also possible to analyze the information locally and then transmit it to a remote network management station. This architecture is called Remote Network Monitoring (RMON).

SNMP defines five protocol messages called Get-request, Get-next-request, Set-request, Get-response and Trap. For example, with Get-request an SNMP server can request information, such as the number of IP datagrams delivered, from an agent. This is called polling. With Trap, an SNMP agent can also deliver information, such as a parameter exceeding a predefined threshold, without a request from the server. The management server then collects the data for various purposes. For example, the data set used in section 3.1 of this thesis was collected using SNMP polling.

Functional deficiencies of the original SNMP were updated by SNMPv2. SNMPv3 in turn brought along authentication and access control. Advantages of SNMP include wide support among vendors and low processor load. On the other hand, the capabilities of SNMP are limited: For example, it only allows data in scalar format. Therefore other protocols have been suggested to replace SNMP.

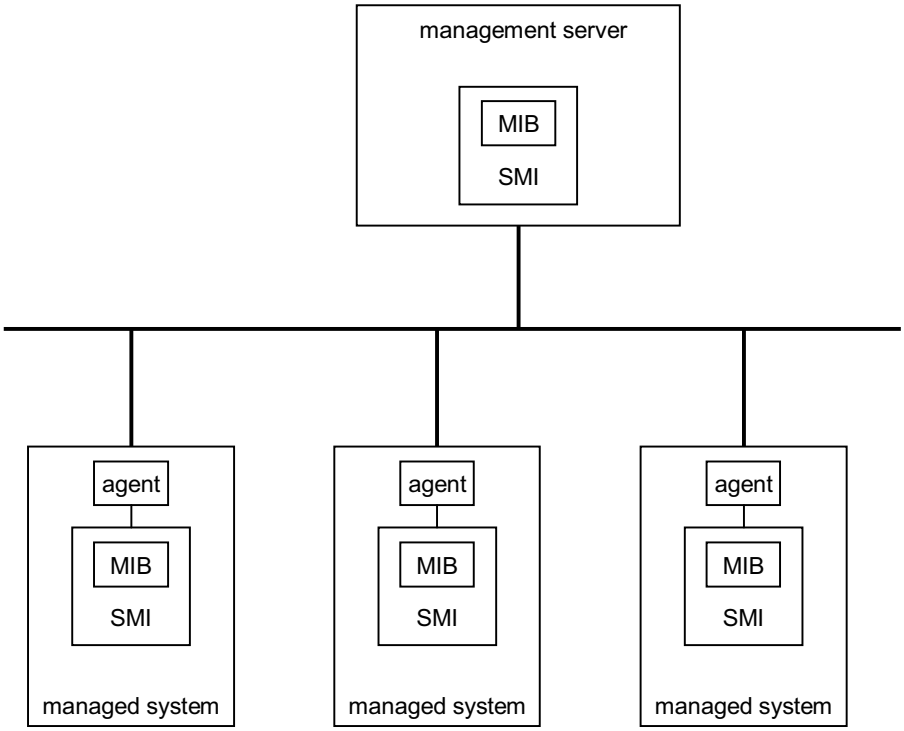

Figure 1.2: SNMP-based network management.

Common Management Information Protocol (CMIP) is a network management protocol based on the ISO/OSI model. It is an object-oriented approach and aims at being flexible, covering a wide range of management needs. The information in CMIP is stored in the form of a Management Information Tree (MIT), where a subtree can, for example, contain or inherit another part of the tree.

Originally, CMIP was expected to replace SNMP in network management. As the popularity of SNMP increased and SNMPv3 was issued, the development of CMIP remained slow. However, together with the higher-level management tool Telecommunications Management Network (TMN), also CMIP is regaining interest.

TMN originated from the need of interoperation between private and proprietary network management systems. It was proposed as early as 1986 by the International Telecommunications Union, Telecommunication Standardization Sector

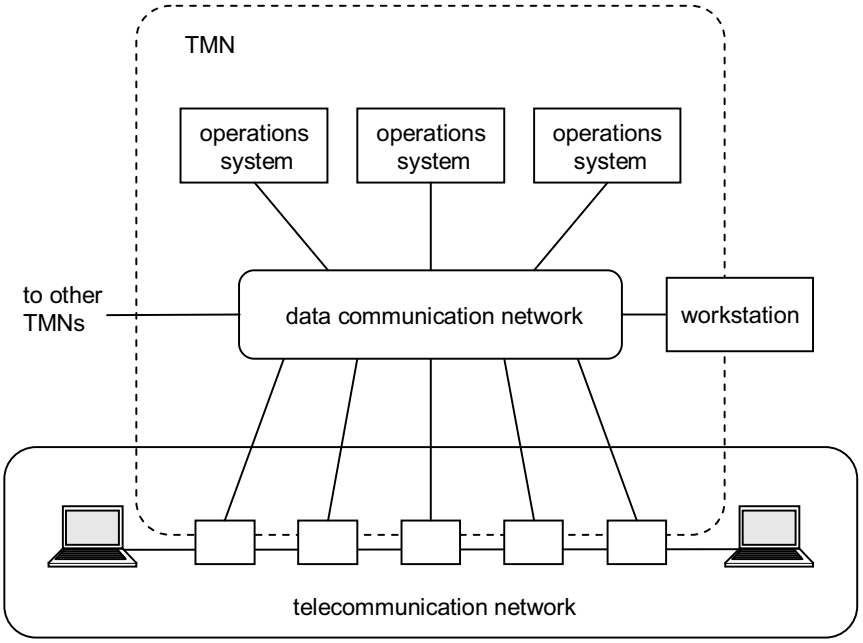

Figure 1.3: The relationship of a telecommunications management network (TMN) to a telecommunication network. [90]

(ITU-T) [90, 108]. The concept has not gained wide acceptance mainly because of its technical complexity and dependence on OSI network management, and the simplicity of SNMP management. Recently, object-oriented software engineering has developed and thus made also TMN more attractive.

The TMN model makes a logical difference between telecommunication and data communication networks (Figure 1.3). The telecommunication network is the network to be managed, whereas the data communication network is a logically separate network that may use parts of the telecommunication network to provide its communications [90]. The operation systems in the upper part of Figure 1.3 execute the OAMP functions. Network management system, traffic measurement system and trunk test system are examples of operation systems [108].

The TMN functionality arises from function blocks, functions and reference points. The recommendation [90] defines five function blocks: operations systems, net-

work element, mediation, workstation, and Q adapter. Each function block contains a set of functions. A reference point is an interface for information exchange between function blocks.

While TMN is a framework for connecting and interoperating data communication networks, it often uses management protocols defined by CMIP. Furthermore, large commercial network management systems usually support all protocols mentioned in this section as well as possible proprietary ones. Thus the data available from a system does not necessarily depend on the underlying protocols. For example, the data used in Chapters 4–6 were collected using NetFlow, which is Cisco's proprietary protocol.

1.4 Contributions

This thesis applies some well-known methodology to the wide field of telecommunication technology. Some methods are refined to suit the specific needs of network data analysis, such as discrete data. This can be viewed as a contribution to the methodology. However, the thesis contributes more to the application field, network management. The methods presented add to the knowledge of handling network traffic data with statistical methods.

One of the guidelines of the thesis is the use of real network data. Most of the chapters include case examples with traces collected from a large campus network, a test environment, and publicly available sources.

The following list introduces the main contributions of this thesis together with references to the respective chapters.

- As a result of the large amount of data, the information is often compressed to histogram format. Chapter 4 presents a method for detecting changes in histogram data.
- Traditional goodness-of-fit tests suit poorly to heavy-tailed distributions. Chapter 5 studies several test statistics with Monte Carlo simulation and

reveals that very simple statistics outperform the more complicated ones presented in the literature.

- Chapter 6 conducts another Monte Carlo study to find an appropriate sample size for heavy-tailed distributions. Instead of the classical concern of samples being big enough, here the sample should not be too big.

- According to some sources, heavy-tailed distributions possess a feature that is useful in prediction. Chapter 7 examines this feature and discovers that it is of little practical importance.

1.5 Structure

The statistical methods used in this thesis rest on firm mathematical theory. Since some elements of the theory appear in several chapters and phases, the necessary theory is gathered together and presented first in Chapter 2. The reader is expected to be familiar with the fundamentals of mathematical statistics, so concepts like normal distribution or density function are not studied very thoroughly. Still, plenty of important details are covered to help understanding the substance of the thesis. The two major topics of Chapter 2 are heavy-tailed distributions and statistical testing.

Chapter 3 serves as an introduction to statistical testing and real-life data. It studies two data sets with a basic Anderson-Darling test and a specific method called censoring. Although the samples do not follow a single analytical distribution, the meaning of the results is worth discussing.

Network traffic data is fairly often discrete in amplitude. Byte and packet counts are common measurements, or round-off errors may cause the discreteness of the data. Sometimes the network management system gives the data as a histogram, which is in essence a discrete random variable. Goodness-of-fit tests for discrete random variables are reviewed in Chapter 4, and a method for detecting changes in histograms is presented.

Chapter 5 searches for the answer to a classical question: What is the most suit-

able goodness-of-fit test for a certain application? Monte Carlo simulation is used to study several test statistics for heavy-tailed distributions presented in the literature. The result is somewhat surprising: the most powerful test statistic to detect changes in the traffic mixture is the median. The more complicated statistics exhibit inferior power characteristics when studied over a gradual change in the traffic collected from a real network.

Sample size is a focal problem in statistical testing. However, usually the question is whether the sample is big *enough*. In network traffic data analysis, collecting huge amounts of data is an inherent part of network management, so increasing the sample size is not limited by resources as usually is the case in manufacturing or medicine. But the question of sample sizes is not relieved by massive data storages, it just turns upside down: Is the sample *too* big? Too big a sample contains excessive information and exposes irrelevant details to the test. This problem is addressed in Chapter 6 by conducting a power study with a wide range of sample sizes from real traffic.

Theoretically, heavy-tailed distributions lend themselves to prediction. This feature however has been neither examined nor applied before. Chapter 7 starts from the definition of a heavy-tailed distribution, induces some features related to prediction and discusses whether the heavy-tailedness really helps in predicting. A simulation example is provided.

Finally, Chapter 8 wraps up the thesis and discusses its significance as well as its weaknesses.

Chapter 2

Mathematical methods

This thesis uses a variety of statistical methods and concepts ranging from simple distribution functions to specific goodness-of-fit tests. As the main application area is telecommunication networks, the reader is not expected to have thorough acquaintance with mathematical statistics. Yet, understanding the meaning of the work requires some knowledge of the methodology as well. This chapter lays a foundation for following the analysis in the rest of the thesis. The topics of this chapter appear in many of the following chapters; in addition, each chapter introduces some more specific theory on its own.

The reader should have a basic knowledge of calculus. Furthermore, a thesis is never a textbook, so this chapter alone is not sufficient for gaining an understanding of statistics. Some fundamental concepts, such as the cumulative distribution function, are rehearsed, but further information can be obtained from, for example, [83, 35, 100].

Different sources have different approaches to the topics of the field. This chapter combines its own approach, leaves out some aspects, such as the most familiar distributions, and emphasizes certain details important to the sequel.

The chapter starts off by defining a random variable. Estimating density and distribution functions have an essential role, since a majority of the analysis concentrates on samples of real network data. Tools for heavy-tailed distributions

receive particular attention. Finally, one of the cornerstones of this thesis, statistical testing, is reviewed. The principles of statistical testing are introduced in a way slightly different from most textbooks, and the subclass goodness-of-fit tests is premised on these principles.

2.1 Random variables

A *random variable* is a real-valued function X defined on a sample space [35]. Depending on the sample space, the variable is either continuous or discrete. Network traffic data contain both types; as a matter of fact, sometimes the limit between continuous and discrete data is unclear. Most data available from a telecommunication network are continuous — durations, for example — but also inherently discrete data, such as packet counts, exist. Chapters 3 and 4 discuss continuous and discrete variables and the sometimes vague difference between them.

Order statistics are often needed in statistic calculations. $X_{(i)}$ denotes the ith order statistic of a sample from X so that $X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$. With continuous variables, equalities may be omitted from the ordering because two realizations of a continuous random variable cannot be equal.

2.2 Density and distribution functions

The *cumulative distribution function* (cdf) of the random variable X is the probability[1]

$$F_X(x) = \Pr\{X \leqslant x\}. \tag{2.1}$$

[1] A capital P is often used to denote probability. Since P however has two other meanings in this thesis — P-value and $P(\theta)$ for power — probability is denoted by $\Pr\{\cdot\}$.

The subscript may be omitted from F_X when there is no risk of confusion.

The cdf associates with the distribution it represents, so it is common to speak of "the distribution F", where F is a cdf. Its derivative, *probability density function* (pdf)

$$f_X(x) = \frac{dF_X(x)}{dx}, \tag{2.2}$$

serves as an illustrative graphical presentation of a distribution. One of the most familiar applications is the pdf of the normal distribution, the famous Gauss curve.

From the definitions follows that the probability of X being in the interval $(a,b]$ is

$$\Pr\{a < X \leqslant b\} = F_X(b) - F_X(a) = \int_a^b f_X(x)dx. \tag{2.3}$$

If $F_X(x)$ is continuous and $a = b$, the value of the integral in (2.3) is zero. Thus, the probability of a single value in a continuous distribution is zero [104]. If the distribution is discrete, the cdf becomes a sum:

$$F_X(x) = \sum_{x_j \leqslant x} \Pr\{X = x_j\} = \sum_{x_j \leqslant x} f(x_j), \tag{2.4}$$

where x_j $(j = 1, \ldots, c)$ are the possible values X can take on.

The terminology involved with the two functions above is sometimes ambiguous. What was named cumulative distribution function here is sometimes referred to as cumulative density function or just distribution function. Probability density function is sometimes introduced as density function or probability function. Some authors prefer calling the pdf of a discrete variable the probability mass function.

2.2.1 Distribution estimation

In practical data analysis, the cdf and pdf are seldom known but have to be estimated from a sample of data. An *empirical cumulative distribution function* (ecdf) estimates the cdf, and a *histogram* serves as an estimate of the pdf. Both are briefly introduced here.

Let $X_1, \ldots, X_n$ be realizations of independent, identically distributed variables. The ecdf of the sample is

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{X_i \leqslant x}, \tag{2.5}$$

where 1_a gets the value 1 when the Boolean operation a is true and 0 otherwise. In words, the ecdf is the proportion of observations less than or equal to x. F_n denotes an ecdf of n observations. Figure 2.1 shows an example ecdf of a relatively small sample, thus the stepwise shape of the curve is clearly visible.

For a histogram, one has to attach a partition of the sample space first. Let $x_0 < x_1 < \ldots < x_c$ be a partition with equal bin widths: $x_j - x_{j-1} = h \quad \forall\, j = 1, \ldots, c$. Now the histogram is the number of observations falling into each bin: [93]

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} 1_{x_{j-1} \leqslant X_i < x_j}, \tag{2.6}$$

where j is chosen such that $x_{j-1} \leqslant x < x_j$, $j = 1, \ldots, c$.

The formula in (2.6) yields a stepwise function that is an estimate of the pdf in the sense that its integral is 1. Because the histogram is most often used for graphical representations, the scaling to unit area is not at all necessary. Instead, $nh\, f_n(x)$ is often used; thus the height of each bar in the graph represents the number of observations in the bin (Figure 2.2). This convention is adopted also in this thesis, and the graph with the division by nh omitted is still called a histogram.

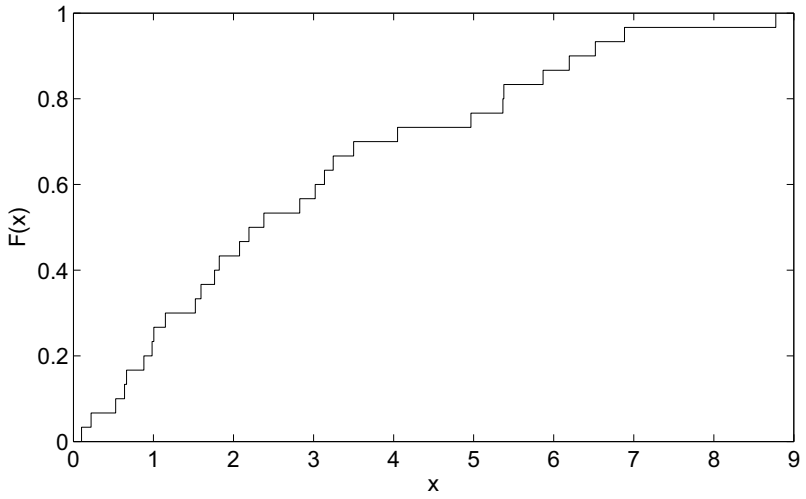

Figure 2.1: Empirical cumulative distribution function of a sample obtained from the exponential distribution ($n = 30$).

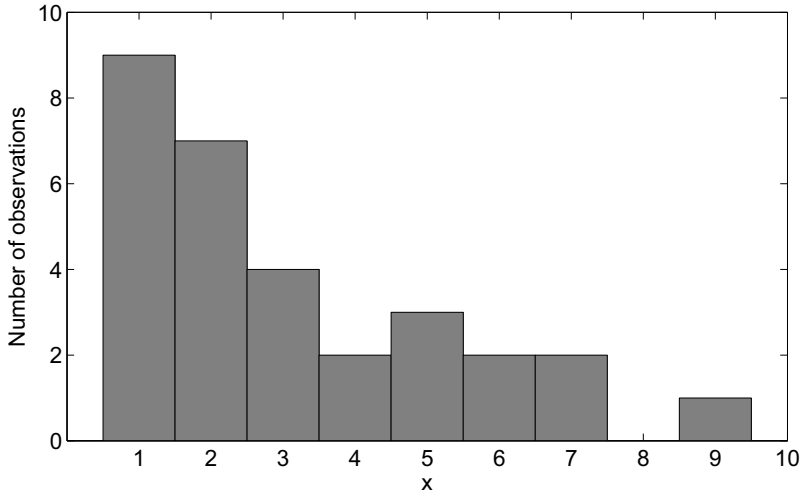

Figure 2.2: Histogram of the sample in Figure 2.1.

Because the shape of the function depends strongly on the partition, the choice of the partition is a subject of a whole discipline [115, 98]. The bin borders need not even be equally spaced, an unequal spacing may lead to a better pdf estimate [33]. Yet, this thesis does not pay very much attention to the partition but chooses bins of equal width more or less intuitively.

2.2.2 Exponential distribution

The exponential distribution is one of the most common distributions in network traffic data analysis. It is considered general knowledge but yet introduced here since some of the chapters in this thesis make use of the exponential distribution. Its cdf is

$$F_{\exp}(x) = 1 - \mathrm{e}^{-\frac{x}{\mu}}, \tag{2.7}$$

and pdf

$$f_{\exp}(x) = \frac{1}{\mu}\,\mathrm{e}^{-\frac{x}{\mu}} \tag{2.8}$$

where the parameter $\mu > 0$ is also the mean of the distribution.

2.2.3 Heavy-tailed distributions

Heavy-tailedness is a property commonly attached to a distribution in network traffic data analysis. A *heavy-tailed distribution* has a right tail that decays as a power-law function:

$$F(x) \sim 1 - kx^{-a}, \tag{2.9}$$

where $k > 0$ is constant and $a > 0$ is a parameter called the tail index. The notation $f(x) \sim g(x)$ means $\lim_{x\to\infty} f(x)/g(x) = 1$, so the power-law shape shows at large values of x. [27] In some texts, this definition is referred to as asymptotical heavy-tailedness.

The Pareto distribution is the simplest heavy-tailed distribution, it therefore appears frequently as a network data model. The Pareto has the cdf

$$F_{\text{Par}}(x) = 1 - \left(\frac{b}{x}\right)^a \tag{2.10}$$

and pdf

$$f_{\text{Par}}(x) = \frac{ab^a}{x^{a+1}}, \tag{2.11}$$

where $x \geqslant b$. The parameter $b > 0$ is often called the location parameter.

Visualization of a heavy-tailed distribution calls for some special methods. The peculiar shape does not visualize well on a linear scale (see Figure 2.3), thus double logarithmic graphs are useful. It has become practice [26, 29, 48] to plot *complementary cumulative distribution functions* (ccdf) $\overline{F}(x) = 1 - F(x)$ instead of ordinary cdf's. In this way, the tail shows up particularly well on the loglog scale (Figure 2.4). Actually, the word "empirical" should prefix also ccdf, since the graph is always plotted based on a sample.

2.2.4 Self-similarity

Consider a stochastic process $Y(t)$ in continuous time. $Y(t)$ is self-similar if

$$Y(t) =_d c^{-H} Y(ct) \tag{2.12}$$

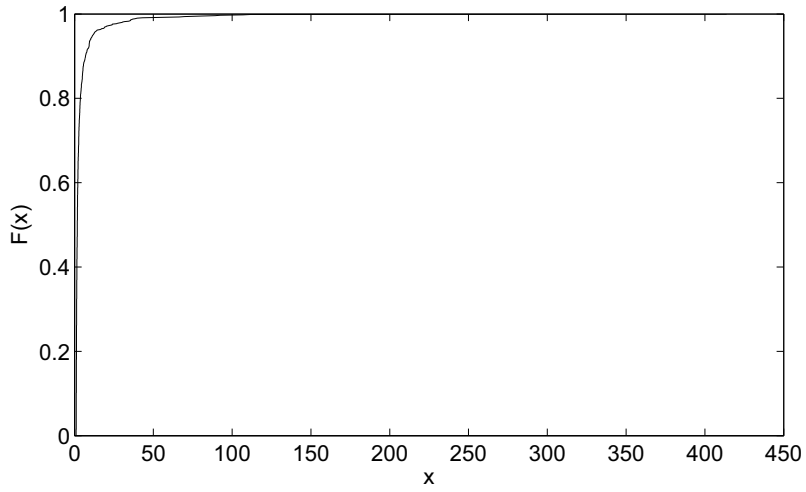

Figure 2.3: Empirical cdf of a sample obtained from the Pareto distribution ($n = 1000$).

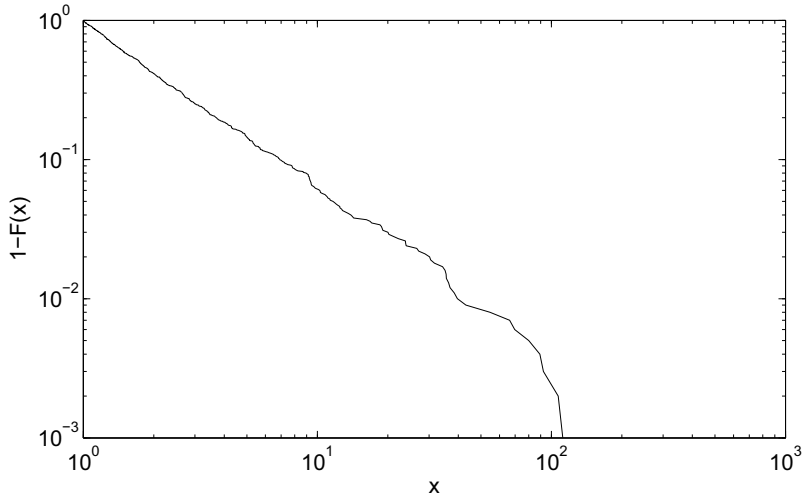

Figure 2.4: Complementary cdf of the sample in Figure 2.3. The theoretical distribution follows a linear slope on a double logarithmic scale, but a sample of finite size has a finite tail.

for any $c > 0$, where $=_d$ stands for equality in distributions. H is called the self-similarity parameter or Hurst parameter after Joseph Hurst, the famous hydrologist who discovered self-similarity in the level of the Nile river. [15]

For self-similar processes, a transformation c in the process time scale is statistically equivalent to a transformation c^{-H} in the amplitude scale [74]. Thus, definition (2.12) manifests itself as a visual resemblance on different time scales.

Taqqu *et al.* [111] proved the connection between heavy-tailed distributions and self-similarity in traffic modelling: Consider traffic sources that are either sending something (on) or idle (off). If the distributions of the on-periods are heavy-tailed with tail index a, aggregating such on/off-sources yields a self-similar process with $H = (3-a)/2$. Hence, in the context of network traffic coming from a large number of sources, heavy-tailed distributions produce self-similarity.

The Hurst parameter takes values in the range $(0,1)$. When $H > \frac{1}{2}$, the process is said to be *persistent*, that is, a value above the mean is probably followed by another value above the mean, and vice versa. This property causes bursts with relatively long periods of high or low values. In *antipersistent* processes, $H < \frac{1}{2}$ and a value below the mean is likely to follow a high value. [21] As bursts are a problem in telecommunication networks, persistent processes are of more interest here. In the light of the above-mentioned connection between self-similarity and heavy-tailed distributions, the range $\frac{1}{2} < H < 1$ can be expressed as $1 < a < 2$ for the tail index.

2.2.5 Long-range dependence

Long-range dependence (LRD) is another term closely related to self-similarity and heavy-tailed distributions. It is not as focal in this thesis as the other two, yet a brief description will be given here.

LRD is essentially a synonym to an autocorrelation function decaying hyperbolically rather than exponentially. The autocorrelation function $r(k)$ of a self-similar process with Hurst parameter H is of the form

$$r(k) \sim H(2H-1)k^{2H-2}, \quad (k \to \infty) \tag{2.13}$$

which implies $r(k) \sim ck^{-\beta}$ when $\frac{1}{2} < H < 1$; β and c are constants such that $0 < \beta < 1$ and $c > 0$. [84] If $H = \frac{1}{2}$, then $r(k) = 0$ and the process becomes uncorrelated. [15]

Heavy-tailedness and LRD also have an intuitive connection. In a stochastic process with long-range dependence, process values that have occurred a long time ago still contribute to the present value. This has an obvious relation to the heavy tail of the cdf: Think of network nodes with activity times coming from a heavy-tailed distribution. As extremely long active periods occur occasionally, these long periods make the present behaviour depend on the situation an extremely long time ago.

2.3 Statistical testing

Statistical testing, or hypothesis testing, is a formal way to estimate the truth value of a hypothesis regarding some random phenomenon. The hypothesis formulates an assumption about the phenomenon, for example:

- Two traffic traces measured from different sources have equal variances.
- The arrival times of the GSM calls in a cell follow an exponential distribution.
- The average delay experienced by an end user is less than the limit that was agreed upon in the service level agreement.

The *null hypothesis* H_0 is supposedly true unless sufficient evidence against it is found, as in the above examples. If H_0 is rejected, the *alternative hypothesis* is accepted. The alternative hypothesis is often of the form "not H_0"; for example, the traces have unequal variances.

Strictly speaking, the hypotheses are not part of the statistical test. Rather, they should be decided before applying the test so that the test procedure cannot affect the hypotheses.

The most important part of the test itself is a test statistic. The following generic procedure describes the course of a statistical test [44]:

1. Design a null hypothesis and an alternative hypothesis to test with data $\mathbf{X} = (X_1, \ldots, X_n)$. Select a test statistic T. These choices also decide the direction of the test: A one-tailed test tests for deviations into one direction only from the null hypothesis, while a two-tailed test detects changes into either direction simultaneously.

2. Derive $F_T(x)$, T's own cdf, either from statistical properties of T or numerically.

3. From data $\mathbf{X}$, calculate T^*, the value of the test statistic for this data.

4. Calculate the P-value of T^* as $P = F_T(T^*) = \Pr\{T \leqslant T^*\}$. The P-value is a measure of the probability of getting a T^* as rare as the one that actually occurred, provided that the null hypothesis is true [104]. A P-value close to either 0 or 1 indicates a questionably rare value of T^*.

5. Using a predefined significance α, define the confidence region of P. If a one-tailed test is to be performed, the confidence region is either $\alpha \leqslant P \leqslant 1$ (test of left tail) or $0 \leqslant P \leqslant 1 - \alpha$ (test of right tail). For a two-tailed test, the confidence region is either the symmetric $\frac{1}{2}\alpha \leqslant P \leqslant 1 - \frac{1}{2}\alpha$ or an asymmetric region of the same length $1 - \alpha$.

6. If the P-value is outside the confidence region, reject the null hypothesis in favour of the alternative hypothesis.

The above procedure contains generalizations and ignores many details of testing. For example, when designing the hypotheses, some assumptions of the data, such as the distribution type, are often made. Furthermore, if the P-value is not available, critical values for T available in many textbooks and computer software can in some cases be used.

Traditionally, terms like "statistically significant" or "highly significant" have been attached to different significance values. Because these terms and fixed α values have their drawbacks, it is increasingly common just to assess the P-value. For example, $P = 0.02$ is more extreme than $P = 0.045$ and arouses more suspicion of an incorrect null hypothesis. If α were 0.05, both would lead to rejection of the null hypothesis.

Example. A service provider has guaranteed that the jitter (variance of delay) in a network connection does not exceed 80 ms^2. To test this guarantee, a sample of 50 delay values is measured. The sample variance is 105 ms^2.

The null hypothesis is "jitter $\leqslant 80$" and the alternative hypothesis "jitter > 80". Assume that the delay is normally distributed. Then the variance of delay can be tested using the statistic

$$T = \frac{s^2(n-1)}{\sigma^2} \tag{2.14}$$

where s^2 is the sample variance and σ^2 the true variance. If the hypothesis is true, T follows the χ^2 distribution with $n-1$ degrees of freedom, where n is the sample size. [44] Since rejection of the null hypothesis requires a large value of T, only the right tail of T's distribution is tested.

Substituting $n = 50$ and $\sigma^2 = 80$ into Equation (2.14) states that $49s^2/80$ be distributed as $\chi^2(49)$. Figure 2.5 illustrates the cdf of this distribution. The figure shows that the P-value of $s^2 = 105$, plotted in the figure as well, is approximately 0.93. Even though the value may sound extreme, its corresponding significance $\alpha = 0.07$ is usually not considered statistically significant. Thus, the hypothesis cannot be rejected in the light of this sample. The service provider has kept its promise. ■

The example contains one major assumption, the normality of the delay distribution. Nevertheless, this assumption is justified. The delay experienced by a single connection is due to several independent reasons, such as congestions in intermediate nodes, varying routes and transit time delays. The central limit theorem states that the distribution of the sum of independent variables, regardless

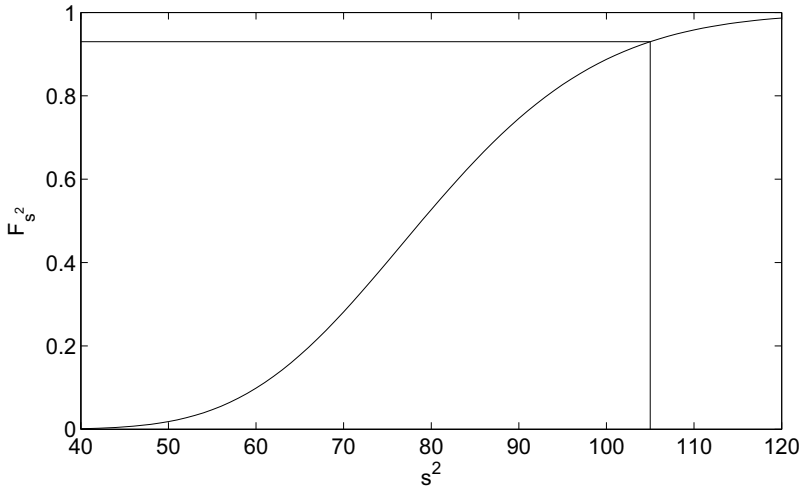

Figure 2.5: Cumulative χ^2 distribution function of $49s^2/80$ with 49 degrees of freedom. The P-value of $s^2 = 105$ is marked in the picture.

of their individual distributions, approaches normal distribution as the number of variables increases [83]. Thus, the delay distribution can be considered approximately normal.

The central limit theorem is one of the reasons why the normal distribution has a salient role in statistics. It is often used in a way similar to the above, to justify the assumption of normality. When stronger evidence for the existence of a certain distribution is needed, goodness-of-fit tests come in. They are reviewed later in this chapter. Additionally, the central limit theorem does not hold for variables with infinite variance, which is the case with many heavy-tailed distributions.

2.3.1 Discrete distributions and statistical testing

In the above, the null hypothesis was rejected if $P < \alpha$. With continuous distributions, it is not possible to get a P-value exactly equal to α, hence the equality in "$\leqslant$" would make no difference. The situation changes when discrete distributions appear.

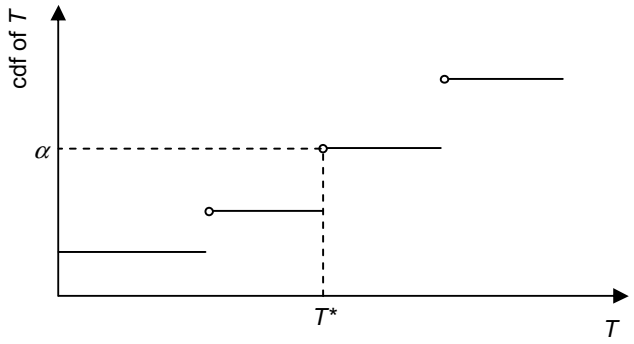

Figure 2.6: Example of a statistical test with a discrete distribution. Solid line: cumulative distribution function of test statistic T, circle: right-continuous point, T^*: sample value of T.

Suppose that X follow a discrete distribution. Now the test statistic T gets discrete values whose probabilities are greater than zero. Accordingly, the P-value can get a value equal to α, and defining the case $P = \alpha$ inside or outside the acceptance region may have a big impact on the behaviour of the test. Figure 2.6 shows a case where T has value T^* whose P-value is equal to α. The rejection condition $P < \alpha$ leads to acceptance of the null hypothesis and thus to a more stringent[2] test than $P \leqslant \alpha$ would. The difference between these two alternatives is proportional to the probability of T being equal to T^*.

The interpretation of the aforementioned situation varies in the literature. For example, Thode [112] uses the criterion $P \leqslant \alpha$ for rejection while Spiegel [104] rejects only if $P < \alpha$. This thesis adopts the latter practice, which yields a more stringent test. This is more natural from the testing point of view: When one is willing to take the risk of incorrectly rejecting a certain proportion α of true null hypotheses, it is safer to make α smaller than bigger.

[2]In the context of statistical testing, the more *stringent* test means the one that has a wider acceptance region, that is, smaller α. The more stringent test requires more evidence against the null hypothesis than the less stringent one.

2.3.2 Monte Carlo simulation

"Derive $F_T(x)$" in step 2 of the above algorithm conceals two main alternatives. If the analytical distribution of the test statistic is known, the exact critical values can be calculated. Most often though, the exact solution is not available, in which case simulation helps solve the problem.

In *Monte Carlo simulation*, sometimes referred to as statistical bootstrapping, random samples with desired statistical properties are generated with random number generators [56]. It is thus possible to construct the cdf of a test statistic by generating a large amount of samples agreeing with the null hypothesis and calculating the test statistic from each of the samples.

The word "simulation" may have different meanings in the field of data analysis. In the above, generating random numbers with certain properties is simulation. Sometimes simulation refers to imitating the behaviour of a real system, such as the GPRS system in Chapter 4. In systems theory, simulation almost invariably means using a mathematical model to simulate the dynamics of a system. Although all these meanings of simulation somehow touch this thesis, the word is reserved for Monte Carlo simulation.

The *inverse transform method* can be used to produce random numbers from a known distribution. Let F be any nondecreasing and right-continuous function with values in the interval $[0,1]$. Then there exists a random variable whose cdf F is. The *generalized inverse function* of $F(x)$ is

$$F^{-1}(u) = \inf\{x \mid F(x) \geqslant u\}, \quad 0 \leqslant u \leqslant 1. \tag{2.15}$$

If F is strictly increasing, then F^{-1} is the ordinary inverse function.

Now if U is a uniform $[0,1]$ random variable, it can be shown that $F^{-1}(U)$ has cdf F. Thus, random numbers from any distribution can be generated by transforming a standard uniform random variable U into $F^{-1}(U)$, as long as F^{-1} is known. [32, 19] See Figure 2.7 for a graphical example of an inverse transform with a truncated cdf.

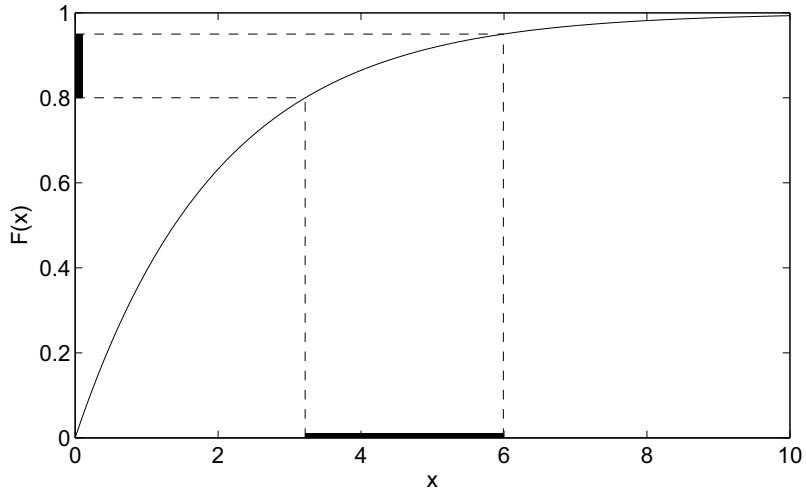

Figure 2.7: Example of an inverse transform. The figure presents a cdf that maps uniform random numbers in the interval $(0.8, 0.95)$ on the y-axis to random numbers in the interval $(3.2, 6.0)$ on the x-axis.

After all, the inverse transform method is based on a uniform random number generator. Virtually all software libraries provide appropriate generators.

The reverse operation, transforming random variable X into uniform random variable through F_X, is called the *probability integral transform* (PIT) [92]. It is commonly utilized in goodness-of-fit testing, since the uniform distribution is easier to test than a more complicated one. The PIT will be used in Chapter 3.

2.3.3 Goodness-of-fit tests

Goodness-of-fit tests are a subclass of statistical tests. Now the null hypothesis H_0 is that a random variable X comes from the distribution $F(x, \theta)$, where θ stands for a parameter, possibly a parameter vector, of the distribution. The alternative hypothesis is commonly just the opposite: x does not follow the distribution $F(x, \theta)$. Two cases are usually distinguished: θ is either known or unknown.

If $\theta = \theta_0$ is known, the distribution $F(x, \theta)$ is completely specified and H_0 is called

a *simple* hypothesis. If θ is not known, $F(x,\theta)$ becomes a family of distributions instead of a completely specified one, and H_0 turns into a *composite* hypothesis. The unknown parameter (vector) then has to be estimated from the sample. [112, 31]

χ^2 test

Pearson's chi-squared (χ^2) test is the classical choice for discrete variable goodness-of-fit testing. Suppose the null hypothesis specifies the probability p_j that a random variable X gets the value a_j ($j = 1,\ldots,c$, where c is the number of possible values of X). Now the test statistic

$$\chi^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j}, \tag{2.16}$$

is χ^2 distributed with $c-1$ degrees of freedom when the sample size is large. O_j and E_j are the observed and expected frequencies of a_j, respectively. The observed frequencies are simply counted from the sample; the expected frequencies can be calculated as $E_j = np_j$, where n is the sample size. Thus, the null hypothesis can be tested with critical values obtained from the χ^2 distribution. [83]

However, the χ^2 test has several drawbacks. It is not suitable for small samples because the test statistic follows the χ^2 distribution only asymptotically. Furthermore, the test does not use the order information of the values, that is, the statistic gets the same value irrespective of the order in which the values are indexed. This makes the test inappropriate for continuous variables and also for discrete variables with natural ordering. Kolmogorov-Smirnov test and its variants suit better to these cases.

Kolmogorov-Smirnov test

Perhaps the most widely used goodness-of-fit test is the Kolmogorov-Smirnov (K-S) test [106]. The K-S statistic is the maximum discrepancy between the hypothesized distribution and the ecdf, thus it is sometimes categorized as a supremum statistic. Most statistical computer programs as well as older textbooks provide tables of critical values for the K-S statistic.

The Kolmogorov-Smirnov test is often commended as distribution free. However, its usefulness in practical applications is limited since it is accurate only in the fully specified case. As soon as parameters have to be estimated from the data, the K-S critical values are no longer valid. Several advanced versions have been developed to compensate for this deficiency. One of them is the Anderson-Darling (A-D) test [8], that will be used in some of the cases of this thesis.

Anderson-Darling test

The A-D statistic A^2 belongs to the class of quadratic statistics, where the squared discrepancy $(F_n(x) - F(x))^2$ is weighted with different weighting functions. The definition of the A-D statistic is

$$A^2 = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 [F(x)(1 - F(x))]^{-1} \, dF(x) \tag{2.17}$$

and the computing formula for a sample

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1)[\ln X_{(i)} + \ln(1 - X_{(n+1-i)})], \tag{2.18}$$

where $X_{(i)},\ i = 1, \ldots, n$ denote the order statistics of the sample. Tables of critical values for the A-D test have been published by Stephens [106].

2.3.4 Power

If a true null hypothesis is rejected on the basis of a statistical test, a *type I error* is committed. The significance α defines the probability of a type I error. On the other hand, if the null hypothesis is really false but gets accepted, a *type II error* occurs. The probability $\beta(\theta)$ of a type II error is a function of θ, the parameter (vector) of the distribution.

The *power* $P(\theta)$ of a statistical test is α when H_0 is true and $1-\beta(\theta)$ when H_0 is false [100]. Ideally, the power should be 0 when H_0 is true and 1 when H_0 is false. If the null hypothesis is a composite one and specifies a region, such as $\theta < \theta_0$, the ideal power curve would step from 0 to 1 when θ exceeds θ_0. Self-evidently, this is not possible. The test statistic, significance α, and sample size among other choices made in the design phase all affect the power, which is always a compromise. Figure 5.2 shows some examples of power curves, and Chapters 5 and 6 discuss the power of various tests.

Chapter 3

Goodness-of-fit tests and network traffic data

Goodness-of-fit tests provide a means of recognizing known distributions from network traffic data. In a sense they constitute an "official" means because of their strong mathematical background with asymptotic distributions, confidence limits, and critical values. Goodness-of-fit methods also contribute to change detection: If one sample fits a distribution and the other does not, inevitably a change has occurred between them.

Paxson [86] criticized network traffic studies for omitting goodness-of-fit tests and also other statistical methods. Quite a few years have elapsed since Paxson's statement, but goodness-of-fit tests still play a minor role in network traffic analysis. Manikopoulos and Papavassiliou [73] used a Kolmogorov-Smirnov test as a similarity measure in conjunction with a neural network classifier. Thottan and Ji [113] made use of a generalized likelihood test for the changes in the parameters of an autoregressive model. Ye *et al.* [120] found the classical chi-squared test to be more appropriate for intrusion detection than another test that takes signal correlations into account. More examples exist, but none of these studies applied actual goodness-of-fit tests to network traffic.

This chapter tries to find the reason why goodness-of-fit tests are not among the most popular methods in network traffic analysis. Perhaps the required mathemat-

ical background is hard to understand or apply to practical cases, or perhaps there just are no known distributions in the network data. No complete answers will be given, the findings rather add to the knowledge of the adequacy of some goodness-of-fit tests. The aim is not only to view the data from the testing perspective but also to view goodness-of-fit tests from the network traffic data perspective.

The study is endorsed by two real-world data sets, one of which comes from a local campus network and the other from a publicly available source. Although heavy-tailed distributions receive a lot of attention both in the literature and this thesis, also other distributions are noticed. Some introduction is provided also to the topics of Chapter 4, discrete goodness-of-fit testing, and Chapter 6, sample size study.

3.1 Normal or lognormal distribution

The normal distribution is the best-known and also the most studied distribution. One of the reasons for this status is the central limit theorem, according to which the average of any random variables is asymptotically normally distributed.

Strictly speaking, there are hardly any true normal distributions in real life. For example, the height of men is often considered normally distributed. But in theory, the normal distribution has support over the real line ensity is always positive, in other words there is neither lower nor upper limit for normally distributed values. However, real-life data most often have an inherent limit, the height for example cannot be negative.

The above example speaks for tolerance in goodness-of-fit testing. Even though the data set seldom fits perfectly into a known distribution, the question is how well it does. This tolerance is controlled with significance; the lower the significance α, the more stringent the test is. Another way to affect the tolerance with which the null hypothesis may be rejected is the sample size, which will be discussed briefly in this chapter and more deeply in Chapter 6.

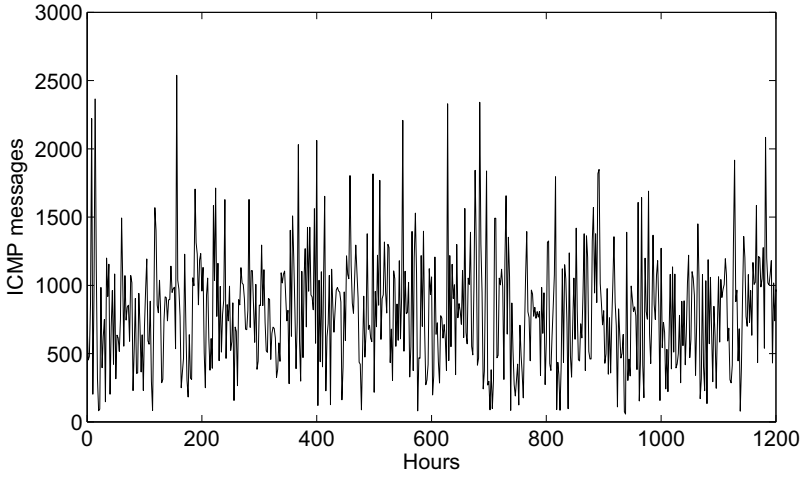

Figure 3.1: Number of outbound ICMP messages from a router.

3.1.1 Visual analysis

The first example data set was collected with SNMP polling from a router in Tampere University of Technology campus network. The data contains traces of various traffic flows with several levels of aggregation. The number of outbound ICMP messages over a period of 50 days (Figure 3.1) represents a typical trace in network data analysis. Message counts were aggregated over 120 minute intervals, and the maximum of each interval brought one observation into the data set, which resulted in 618 observations altogether. The graph in Figure 3.1 presents the data as a time series, while in Figure 3.2 a histogram shows how many observations fall into each 100 units wide bin.

The time series looks like random noise, a sample of a normally distributed random variable. An experienced data analyst may however notice some skewness in the time series already, and the histogram strengthens this suspicion. The distribution most obviously has a positive skew, though also the shape of the normal distribution is clearly visible. The histogram is used here quite carelessly as a tool of visual analysis; the same could have been done more formally with an empirical cumulative distribution function.

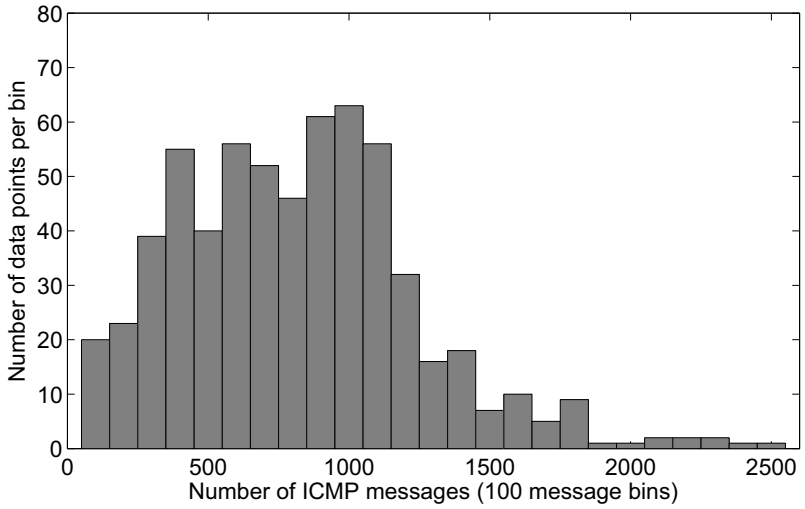

Figure 3.2: Histogram of the data in Figure 3.1 with 100 message bins.

The shape of the histogram suggests a lognormal distribution. To justify the guess of normality or lognormality, Figure 3.3 shows the ecdf together with the theoretical cdf's of the distributions in question. Naturally, neither of them provides a perfect fit. Perhaps the normal distribution seems to fit the data better in the right tail and the lognormal distribution in the left tail.

Notice that the cdf's in Figure 3.3 cannot be used in the Kolmogorov-Smirnov sense (section 2.3.3). The K-S test works only for a completely specified hypothesis, that is, when all parameters of the distribution are known [106]. The cdf's in the figure were plotted using parameters estimated from the data, which ruins the basic idea of the Kolmogorov-Smirnov test.

3.1.2 Anderson-Darling test for normal and lognormal distributions

The Kolmogorov-Smirnov test does not allow estimation of distribution parameters from the data. Since in practical applications the parameters are rarely known beforehand, more powerful tools are needed. The Anderson-Darling (A-D) test

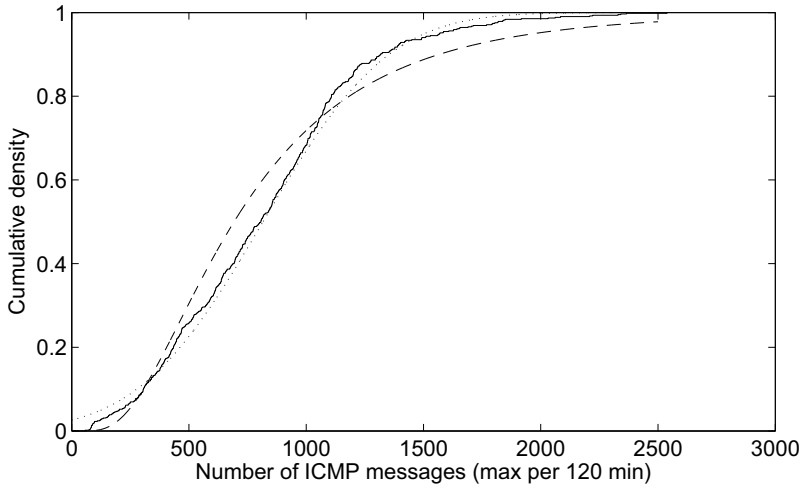

Figure 3.3: Empirical cdf of the data presented in Fig. 3.1 (solid line), and theoretical cdf's of a normal (dotted line) and a lognormal (dashed line) distribution.

for normality can be used even if parameters are estimated from the sample, but different critical values apply to the completely specified case and estimated parameters. Stephens [106] suggests a sample size dependent modification to the A-D test statistic A^2 (section 2.3.3) to be used when parameters are estimated:

$$A^{2'} = A^2(1.0 + 0.75/n + 2.25/n^2), \tag{3.1}$$

where n is the sample size. He also provides a table of critical values for $A^{2'}$. The modified statistic $A^{2'}$ is used in this chapter.

The data in Figure 3.1 can now be tested for normality and lognormality. For the normal distribution, the parameters to specify the distribution are the mean μ and the variance σ^2. They must be estimated with their sample counterparts

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.2}$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - m)^2. \tag{3.3}$$

Calculating the probability integral transform (PIT) [92] requires the distribution function of the normal distribution. It does not exist in a closed form, thus a numerical approximation obtained from a computer program must be used.

Now the PIT from normal to uniform distribution can be carried out, and the modified A-D statistic gets the value 2.879. According to [106], the highest critical value (corresponding to significance level 0.005) is 1.159. Thus, the test statistic is extremely rare in its own distribution and the normality hypothesis must be rejected.

How about lognormality? If X is distributed lognormally, then $\ln X$ is distributed normally. The value of the modified A^2 for $\ln X$ is 11.971, which is much further off the critical values than in the normality test. In spite of the slight skewness seen in the histogram (Figure 3.2), the hypothesis of lognormality is rejected.

3.2 Pareto distribution

This section studies the use of the A-D test for heavy-tailed distributions in the light of an example data set. The Pareto distribution is used as a prototype of heavy-tailed distribution because of its simplicity.

3.2.1 Visual analysis

Crovella and Bestavros [26] analyzed WWW traffic traces from Boston University. The data set is available in [41] and described in detail in [30]. This section uses the same data to view the difficulties in using standard goodness-of-fit tests

for real-life data. Henceforth, the data set analyzed by Crovella and Bestavros is called the Boston data.

Figure 3.4 is most similar to a figure in [26]. It contains the complementary cumulative distribution function (ccdf) of file transmission times of the Boston data as well as a linear trend fit to the ccdf. A minor difference at the lower end comes from the exclusion of times smaller than one second, which was stated in the text of [26] but not implemented in the original figure.

A time series representation of the transmission times presented in Figure 3.5 reveals a clear daily variation. The sample is considered stationary in spite of the variation, since the time span of more than a month drowns the daily cycle. Shifting the time span would not change the distribution of the data.

Fitting a linear trend to a cumulative distribution function, as is done in Figure 3.4, is commonly regarded as a way of finding the tail index, the exponent of the heavy-tailed distribution. Because of the log-log scale, the slope of the line is inherently the tail index:

$$\frac{d\log\overline{F}_{\mathrm{Par}}(x)}{d\log x} = \frac{d\log(\frac{b}{x})^a}{d\log x} = -a \tag{3.4}$$

The line in Figure 3.4 has slope -1.21 as determined in [26]. The fitting however is not straightforward; several approaches to determining the slope of the line have been proposed [49, 28, 1]. A closely related field is estimation of the Hurst parameter, widespread methods of which include the R/S method and the Whittle estimator [110].

The tail of a theoretical heavy-tailed distribution extends to infinity. In a real-world sample however, the amount of large values is always limited. For example, the empirical distribution in Figure 3.4 decays at approximately 1000 seconds. It therefore suffices to inspect that the tail shows linearity over several orders of magnitude [27]. Crovella and Bestavros [26] phrase this as follows: “The shape of the upper tail on this plot, while not strictly linear, shows only a slight downward trend over almost four orders of magnitude.” Along with this statement, the heavy-tailedness of the distribution is examined next with goodness-of-fit tests.

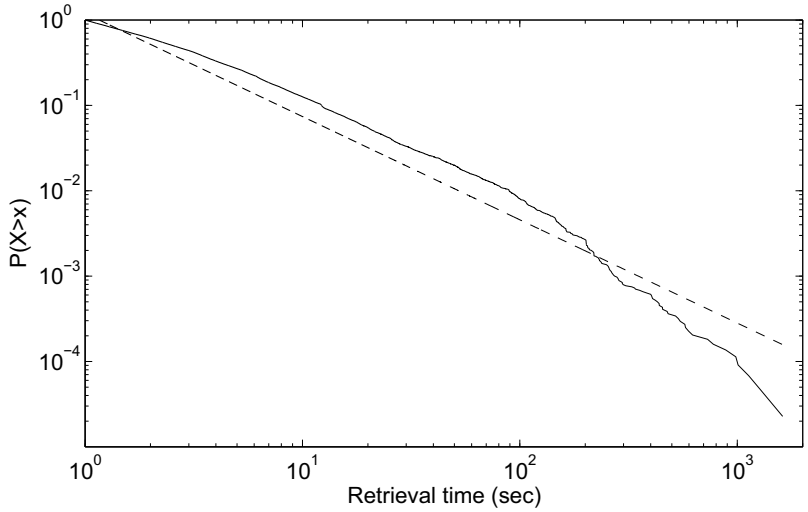

Figure 3.4: Empirical distribution of file retrieval times. Solid line: empirical ccdf. Dashed line: linear fit.

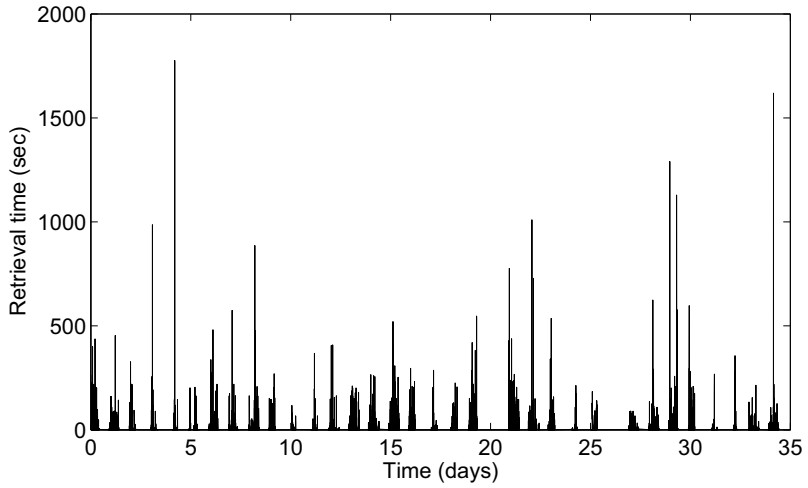

Figure 3.5: File transmission times of Boston data in February 1995.

3.2.2 Anderson-Darling test for censored Pareto distribution

In Figure 3.4 only values greater than 1 second were taken into account. Usually large values in the data interest network designers because extremities cause congestions and thus set conditions to the design. Other, secondary reasons include the discreteness of the data among the small values and even zero values due to measurement inexactness. Additionally, the Anderson-Darling test weights both tails of the distribution [8], which in the case of a heavy-tailed distribution would mean weighting the small bulk values in the left "tail".

In statistical testing, selecting data according to a certain criterion is called *censoring*. When only the (right) tail of a distribution is tested, only values greater than a certain limit are included in the test. This method is known as type 1 left-censoring [77]. Type 1 left-censoring is typical of network data analysis because of the above reasons. Next, a goodness-of-fit test for left-censored Pareto distribution based on [106, 92, 77] is presented and applied to the Boston data.

Let $X_i,\ i = 1,\ldots,n$ be a random sample. The null hypothesis is that X_i comes from the Pareto distribution

$$F_{\mathrm{Par}}(x,a,b) = 1 - \left(\frac{b}{x}\right)^a. \tag{3.5}$$

Substituting $x = e^y$, $a = \frac{1}{\mu}$ and $b = e^\theta$ yields the exponential distribution:

$$F_{\mathrm{Par}}(x,a,b) = 1 - e^{\theta/\mu}e^{-y/\mu} = 1 - e^{\frac{-(y-\theta)}{\mu}} = F_{\exp}(y,\mu,\theta), \tag{3.6}$$

which proves the well-known connection between the Pareto and exponential distributions: If X is Pareto distributed, then $Y = \ln X$ is exponentially distributed.

θ and μ in Equation (3.6) stand for the location and scale parameters of the exponential distribution, respectively. This parametrization can be easily returned

to the more familiar case $\theta = 0$ by setting $Y'_i = Y_i - \theta,\ i = 1,\ldots,n$ or, if θ is unknown, $Y'_{(i)} = Y_{(i+1)} - Y_{(1)},\ i = 1,\ldots,n-1$. Put in words, subtracting the minimum value from an exponentially distributed random sample yields another exponentially distributed random sample. This result can be applied successively r times to give a random sample of size $n-r$. [106] This will be applied next.

Because subtracting r smallest values from an exponential sample does not change the hypothesis of exponential distribution, testing a left-censored sample is straightforward. If the sample is left-censored in the original sense, it is known to contain only values greater than a certain threshold l. Or, as is convenient with heavy-tailed distributions, the complete sample can be reduced by removing all values smaller than l. In both cases, the smallest value of the censored sample can be subtracted, which entails a new sample that hypothetically comes from an exponential distribution with $\theta = 0$.

The test procedure for the left-censored Pareto distribution is now ready. Here it is applied to the sample in Figure 3.4:

1. Transmission times below 1 second were left out of the figure, because large values are of more interest. Hence, censor the sample to the left of $l = 1$. 44063 values remain in the sample.

2. Transform X to $Y = \ln X$.

3. Estimate μ with the sample mean m and use the probability integral transform $Z_i = 1 - e^{-Y_i/m}$.

4. Calculate the A-D statistic according to Equation (2.18) and compare it to critical values given in [106].

5. The value of the statistic is 286.8, which is far beyond even the highest available critical value, 2.24 for significance 0.005. Thus, the hypothesis of Pareto distribution is rejected.

3.2.3 Sample size considerations

The above result reveals an important aspect: sample size. The sample of 44000 observations is huge compared to normal textbook examples and has no chance to pass the goodness-of-fit test in spite of the sample size dependent modification to the A-D test. Using all information in the data set simply reveals even the slightest deviations from the hypothesis.

Real-world data sets never come from a fully defined distribution, they always contain features from several different sources, such as different users, varying conditions and dynamic phenomena. There are tools for handling these features, for example mixed distributions or time series models. Nevertheless, the complicated distributions hinder the use of classical goodness-of-fit tests together with large samples.

In terms of test power (the probability of rejecting a truly false null hypothesis, see Chapter 2), increasing sample size greatly increases the power of the test. Even though exploiting all the information in the large sample may sound like a good idea, excessive power is not a desirable feature of a statistical test. Sample sizes and statistical power will be studied in Chapter 6. This section approaches sample sizes in the light of the above results and thus sets up the discussion of Chapter 6.

Table 3.1 gives some insight into the effect of sample size. It contains the A-D test statistics and critical values (c.v.) for significance level $\alpha = 0.05$, four sample sizes and four null distributions. The critical values for normal and lognormal distributions are the same because they were tested using the same test, the latter one with the transformation $Y = \ln X$. The same applies to exponential and Pareto distributions.

When the sample size is big, $n = 2500$ or $n = 500$, the A-D test statistics for normal and exponential distributions cannot be calculated, which is marked with "infinity" in the table. This is due to more than one equal smallest values in the sample, which causes log of zero in Equation (2.18). Hence, coarse resolution causes slight discreteness among the smallest values of the sample, although the data set is generally considered continuous (for further discussion on discrete data, see Chapter 4). Also the test statistics for lognormal and Pareto distributions yield

Table 3.1: Test statistics and critical values ($\alpha = 0.05$) of Anderson-Darling tests for some known distributions and different sample sizes.

	Normal	Exponential	Lognormal	Pareto
n	c.v. 0.752	c.v. 1.321	c.v. 0.752	c.v. 1.321
20	2.305	1.304	0.485	0.571
100	15.981	16.035	2.369	0.868
500	infinity	infinity	13.479	2.921
2500	infinity	infinity	64.170	15.089

values significantly larger than the corresponding critical values.

The sample of 100 observations leads to acceptance of the Pareto distribution but rejection of the other hypothesized distributions. This suggests that the Pareto may be the closest candidate distribution of these four, but the conclusion is far from clear. Drawing 100 observations out of 44000 leaves plenty of variation in the sampling, therefore the test result may be totally different for the next 100 observation sample. Actually, all that a statistician can say is that the test statistic exceeds the critical value in no more than 5 % of the cases *if* the null hypothesis is true. Colloquially, if the sample really comes from the Pareto distribution, it is 95 % certain that the test statistic value does not exceed the critical value.

Furthermore, the row of the small sample ($n = 20$) shows that on the grounds of this sample, neither exponential nor lognormal nor Pareto distribution can be rejected! Still, the data does not come from several different distributions but information is lost by picking a small sample. The limited information is not sufficient for rejecting the null hypothesis. This is also a reason why one should speak of "not rejecting" rather than "accepting" the null hypothesis.

A test statistic that is "closer" to a critical value than another test statistic does not provide any information as such. For example, the test statistic for the lognormal distribution (0.485) is closer to its critical value (0.752) than the Pareto test statistic (0.571) to its own critical value (1.321). Yet this does not mean that the sample "more likely" comes from the Pareto distribution. If P-values or some other infor-

mation on the distributions of the test statistics were available, one could conclude which hypothesis is closer to acceptance or rejection.

3.3 Conclusion

The use of goodness-of-fit tests in network traffic data analysis is relatively rare. One reason for this is that these tests do not provide any easy solutions to data mining. The distributions that appear in telecommunication networks are seldom easily uncovered with goodness-of-fit tests. Rather, the traffic contains burstiness over many time scales, as reported in several publications [66, 87, 38].

Thus, the question is not whether the sample positively comes from a known distribution but whether it *resembles* one closely enough. If an approximate analytical distribution can be found, it can be used as a model of the data, which opens plenty of opportunities to mathematical analysis [86]. On the other hand, an analytical model can never reflect all features of the true data.

Chapter 4

Change detection of discrete data

The previous chapter discussed briefly the discreteness of data and decided that the effect of discretization in the Boston data was negligible. In the end, all digital measurements are discrete. Thus the limit whether methods meant for continuous distributions can or cannot be used is imprecise [106].

However, many quantities measured from a telecommunication system are inherently discrete. This may be due to several reasons:

- The quantity being measured is a count, such as the number of packets in a flow or number of new connections per time unit.
- The measurement unit is coarse compared to the magnitude of the quantity. For example, flow durations may be measured with one second accuracy although the average duration is only a few seconds.
- Rounding to discrete values occurs for some other reason. Because the data collection systems involved in network nodes are complex, sometimes the reasons for rounding are not even known.
- The data come in as histograms or contingency tables. These are commonly used for compressing the otherwise large amount of information. A histogram or contingency table is essentially a discrete random variable.

Table 4.1: Relations of various measurement types. '+': possible combination, '–': not possible.

		discrete	continuous
categorical	nominal	+	–
quantitative	ordinal	+	–
	interval	+	+
	ratio	+	+

A discrete random variable only takes values of a countable set, either finite or infinite. A discrete variable can be either quantitative or qualitative [114]. Quantitative variables have meaningful ordering, possibly also meaningful intervals and ratios, while qualitative variables — colours, for example — lack ordering and distance measures. In data analysis, qualitative variables are often called *categorical*. In the following examples, port numbers are a categorical variable and interarrival times a quantitative one.

Another common classification of measurements is the division into nominal, ordinal, interval and ratio scales [89]. Table 4.1 summarizes how this division relates to discrete, continuous, categorical and quantitative measurements.

This chapter starts off by comparing two samples with the χ^2 test, the most commonly used statistical test for discrete data. As was stated in Chapter 2, it has its drawbacks but here the χ^2 test is applied to where it is at its best, a set of categorical data that has no natural ordering. For a set of quantitative data, Monte Carlo simulation together with the Kolmogorov-Smirnov statistic provides a better alternative, which is proved by way of an example with data from a GPRS test arrangement.

The emphasis is now not on goodness-of-fit testing in the purest sense: The tools of this chapter do not search for an analytical distribution that would fit the data. Instead, methods familiar from goodness-of-fit testing are used for change detection together with Monte Carlo simulation.

Table 4.2: Contingency table of port numbers in two samples. O_{ij} = number of observed values j in sample i.

	Port 1	Port 2	$\cdots$	Port c	Total
Sample 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Sample 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
Total	m_1	m_2	$\cdots$	m_c	n

4.1 Comparing samples of discrete distributions

A popular use of the χ^2 test is to test whether two (or more) samples come from the same distribution. The samples are arranged into a contingency table, see Table 4.2 for an example. If the null hypothesis of equal probabilities between samples is valid, the expected frequency of X_{ij}, value j in sample i, is

$$E_{ij} = \frac{n_i m_j}{n} = \frac{\sum_{k=1}^{c} O_{ik} \sum_{k=1}^{2} O_{kj}}{n}, \tag{4.1}$$

where n_i is the size of sample i, m_j the number of observations of value j in both samples, O_{ij} the number of observed values j in sample i, and $n = \sum m_j = \sum n_i$. Now the test statistic can be calculated using Equation (2.16). The procedure can straightforwardly be generalized to more than two samples. [100]

4.1.1 Comparing port number histograms

When a network host connects to another, it uses numbered ports as virtual interfaces. Thus, every connection occurs between two ports. Usually one of the end points is a server and the other one a client. Even if neither one actually operated as a server, one of the end points would first listen to a port and wait for the

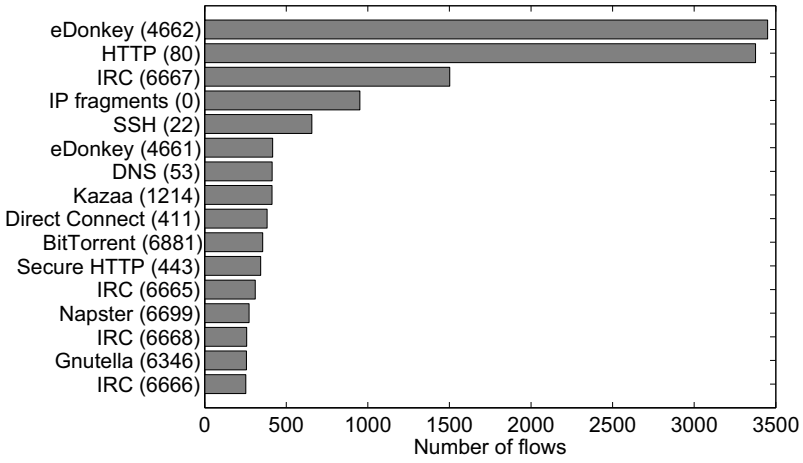

Figure 4.1: Sixteen most frequently appearing destination ports of a data file. The labels on the vertical axis give the names of the protocols associated to each port number.

other one to connect. As the port numbers have to be pre-determined somehow, servers usually use protocol-specific ports. For example, an HTTP (web) server listens to port 80 and a DNS server to port 53. These ports are termed well-known ports, managed by the Internet Assigned Numbers Authority [54]. In addition to the well-known ports, many applications use port numbers by which they can be identified. These registered ports can be found either in IANA's list or various other sources.

The following examples use data of unidirectional network flows collected from Tampere University of Technology campus network with Cisco's NetFlow [24], where each flow contains several details, source and destination ports among them. The collection system divided the flows into 5-minute files, each of which contains approximately 50000 flows. First, one single file is studied mainly for rehearsal, then successive files are compared with each other with the χ^2 test to detect changes.

χ^2 statistics from a single data set

Figure 4.1 presents a histogram of the most frequent destination ports in a data file collected in November 2003. Some applications, such as IRC, appear more than once in the histogram because they use several ports. These several ports of one application could have been merged but it would not have brought any benefit to the analysis. Furthermore, destination port numbers alone do not tell the whole truth of the traffic since for example HTTP uses port 80 as both destination (client request) and source (server response) port. However, for the sake of change detection the bare destination ports will suffice.

In fact, the application names in Figure 4.1 are just guesses, but on the grounds of these guesses ports typical to numerous peer-to-peer applications (Direct Connect, Gnutella, eDonkey, Kazaa, Napster, BitTorrent) seem to dominate the traffic. Some of them have even become obsolete in a couple of years, which manifests the rapid evolution of network traffic. The frequently appearing port 0 is not really a port but NetFlow's way to mark fragments of IP packets.

For the best result, the expected frequencies in the χ^2 statistic should be approximately equal [51]. In this data they are not; Figure 4.1 shows that different applications have drastically different frequencies — the frequencies might even follow a heavy-tailed distribution. If the random variable had a quantitative scale, one could combine consecutive values to achieve more even frequencies. But for a categorical variable, this would require expertise on the applications; for example, HTTP (80) and secure HTTP (443) could be combined.

In spite of the inequal frequencies, the χ^2 statistic works. The results in Figure 4.2 were calculated from 2×100 samples drawn from a population of 40000 flows. The size of each sample was 1000. Port numbers used in this example were the ones presented in Figure 4.1 as well as one category called "others", which was by far the most frequent category. Since there were 17 categories, the test statistic follows the χ^2 distribution with 16 degrees of freedom, as can be seen from the solid and dotted lines in the figure. If the sample pairs had been drawn from different distributions, the solid line would not follow the dotted line.

It is important that the samples be drawn with replacement, that is, an observation

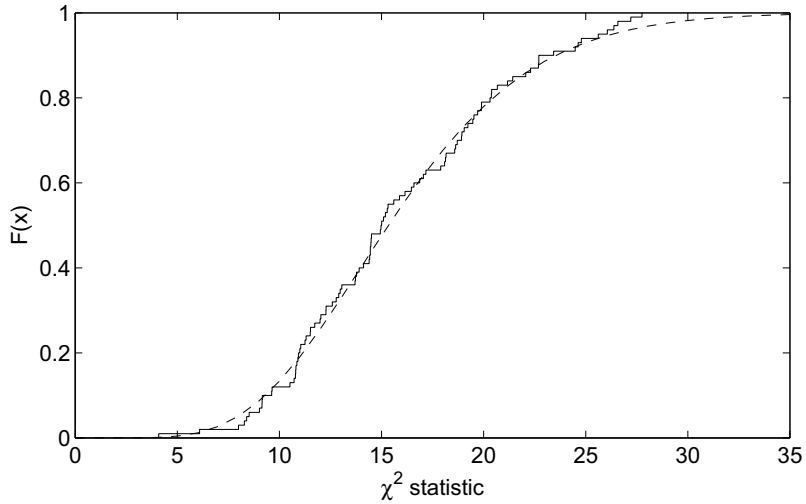

Figure 4.2: Cumulative distribution function of χ^2 statistic. Solid line: empirical cdf calculated from 200 samples. Dashed line: theoretical χ^2 cdf with 16 degrees of freedom.

is "returned" to the population and can thus end up in the sample more than once. If the sample were drawn without replacement, the probability E_i/n of a certain port number would not remain constant. Sampling with replacement may not sound a natural way of sampling network traffic flows: How is a flow "returned" into the process after being measured? However, the usual case is just like the one presented: The data are collected into a file and the analysis is done off-line with the data file. Then the replacement can be easily arranged.

Of course, the effect of the replacement is negligible as long as the sample size is just a few per cent of the population or less. On the other hand, the sample must not be too small: If coincidentally neither one of the compared samples contains any one of the observed port numbers, $E_i = 0$ and there will be a division by zero in Equation (2.16). In other words, the expected frequency may not be zero in any category.

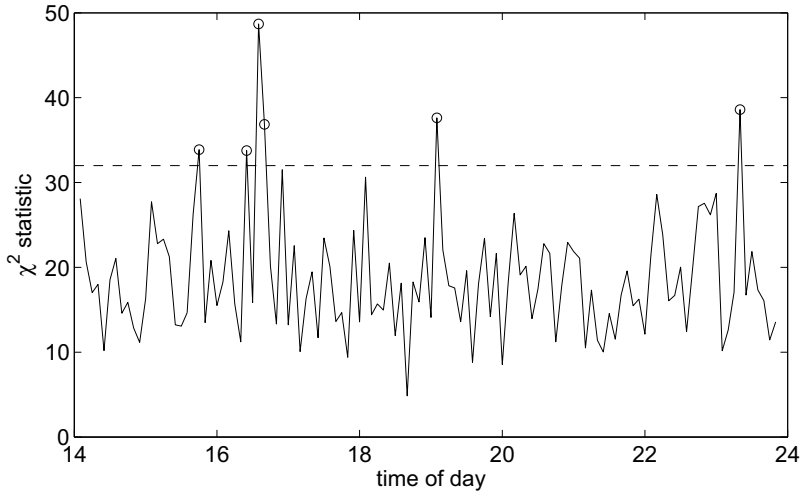

Figure 4.3: Change detection of port distributions with χ^2 test. Solid line: χ^2 statistic of two consecutive 5-minute intervals. Dashed line: 99 % critical value of the corresponding (16 degrees of freedom) χ^2 distribution. Circles: statistics exceeding the critical value.

Comparing two data sets

Recall that the data came in files covering a 5-minute period and tens of thousands of lines. Now, the χ^2 property demonstrated above can be used in change detection of the port histograms. Figure 4.3 illustrates a series of χ^2 statistics calculated from two consecutive data files.

The statistics were calculated using the method introduced above, except that the earlier of the two data files was not sampled but used as a reference as such. The expected frequencies were not obtained from Equation (4.1) but the first line of the contingency table (Table 4.2) got large values and was used directly as expected frequencies. Thus, the risk of getting a zero in the denominator of Equation (2.16) was avoided.

Figure 4.3 also contains a line at the 99 % critical value of the $\chi^2(16)$ distribution. Statistics above the critical value, marked with a circle, indicate a likely change in the port distribution between two consecutive 5-minute intervals. Expressed

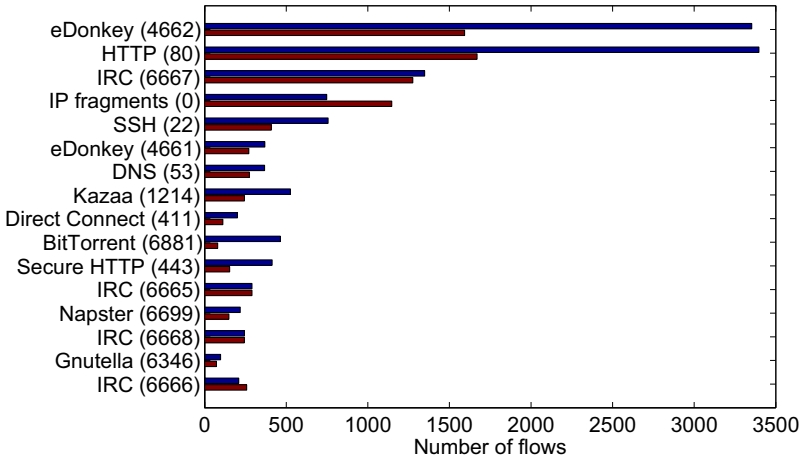

Figure 4.4: Comparison of two consecutive port histograms when the χ^2 statistic is high. Blue bars indicate the first file, red bars the second.

exactly in statistical terms, the chance that the statistic exceeds the critical value is only 1 % if the hypothesis of similar distributions is true.

The highest peak of Figure 4.3 occurs around 16:40 and the lowest one at 18:40 o'clock. Figures 4.4 and 4.5 compare these histogram couples to see if the differences indicated by the statistic are visible. The histogram couple with the high statistic and thus probable discrepancy is in Figure 4.4, while the couple in Figure 4.5 is the one with the low statistic.

In Figure 4.4, one can see radical changes in port frequencies. For some reason the two most popular applications, eDonkey and HTTP, have lost almost halves of their frequencies. Quite clear changes appear also with other applications. In contrast, the bars in Figure 4.5 show much more similarity, there are only slight differences between these two files. On the basis of these graphs it is easy to agree with the χ^2 statistic and reject the hypothesis of similarity in port distributions between the two data files in Figure 4.4. If there were a monitoring system detecting changes in the traffic profile, it should give an alarm at 16:40.

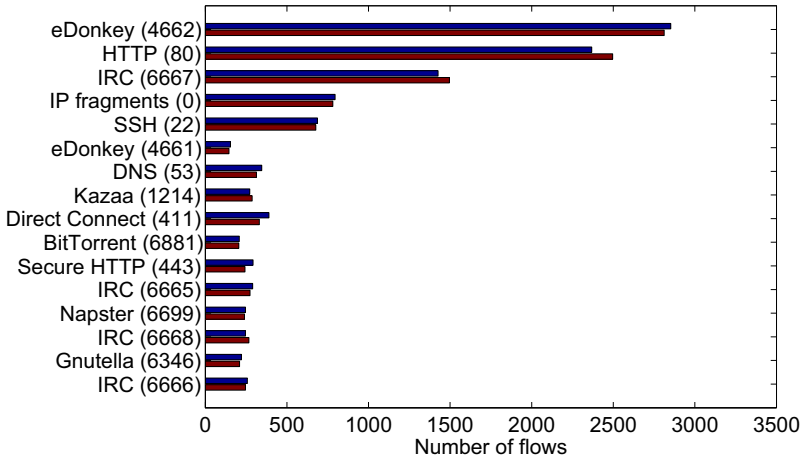

Figure 4.5: Comparison of two consecutive port histograms when the χ^2 statistic is low. Blue bars indicate the first file, red bars the second. Compare with Figure 4.4.

4.2 Change detection of histogram data

Histograms are commonly used in data acquisition for compressing large amounts of data. As the histogram is essentially a density function estimate for the underlying distribution, it provides a quick visual outline of the distribution and its parameters. This section introduces a method for detecting changes when data are available in the form of histograms.

In Chapter 2, the histogram was defined as a piecewise constant function, but in practice a histogram is often given as a contingency table. The table consists of tuples (b_j, m_j), where b_j denotes a bin center and m_j the number of observations in the bin. If the histogram uses bins of equal width, as it usually does, it is fully specified by these tuples.

Let X_i, $i = 1, \ldots, n$ be realizations of independent, identically distributed variables. Presented as a histogram or a contingency table, the information content of the n observations is compressed into c tuples, where c is the number of bins in the histogram. This is equivalent to rounding each observation to nearest bin center b_j, $j = 1, \ldots, c$. Now the sample can be reproduced using the histogram and

the inverse transform method presented in Chapter 2. The reproduced sample is a discretized version of the original one: It has more or less similar statistical properties but observations can only take values b_j. Information is lost in the histogram representation — how much is lost, depends on the number of bins in the histogram.

Testing of histograms has traditionally been a field of the χ^2 test. Advanced variants [79, 51] have been proposed to overcome the deficiencies of the test, yet most authors [51, 106] do not recommend the use of χ^2 tests when the data contains ordering information. Thus, this section rather uses the Kolmogorov-Smirnov (K-S) statistic.

The K-S *test*, as presented in Chapter 2, is not as eligible for discrete distributions as it is for continuous ones, since it not distribution-free even in the case of a fully specified simple null hypothesis [118, 59]. But here the K-S *statistic* is used merely as a measure of fit, while the critical values are sought using Monte Carlo simulation.

Conover [25] showed how the exact critical levels for the K-S statistic for discrete distributions can be found, but the method is reasonable only for relatively small sample sizes ($n \leqslant 30$). Steele [105] strongly encouraged using Monte Carlo methods instead of asymptotic approximations, though his work discussed categorical variables where the values have no natural ordering. Along with the increasing computation power, Monte Carlo simulation seems to be superseding complicated theoretical studies in data analysis practice.

Krishnamurthy *et al.* [61] used the Kullback-Leibler (KL) distance in change detection of network traffic. They used the discrete form of the distance:

$$KL(p||q) = \sum_j p_j \ln \frac{p_j}{q_j}, \tag{4.2}$$

where p_j and q_j are the probabilities of the histograms to be compared, such that $p_j = m_j/n$. The KL-distance is suitable for measuring the divergence between a sample and a reference distribution (discrete or continuous), but comparing two discrete samples is less appropriate for the following reasons.

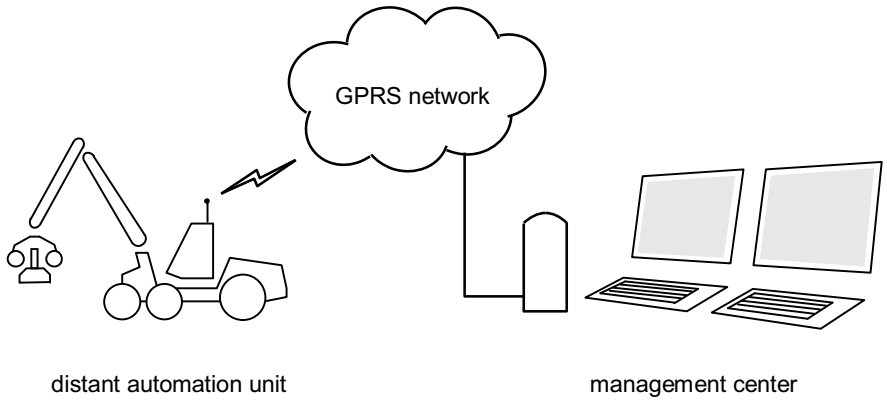

Figure 4.6: A distant production unit that contains stand-alone automation and sends monitoring data over GPRS network.

The disadvantage of the KL-distance in comparing two samples is the possibility of zero values. If $q_j = 0$, $KL(p||q)$ tends to infinity. Krishnamurthy *et al.* bypassed the problem "with a correction for the case when a bin count is zero." However, this may distort the result. The distance is weighted by p_i so that more weight is given to bins where p_i is large [18]; making corrections when q_i is zero but p_i is large inevitably conflicts this intention. Furthermore, the KL-distance is asymmetric, that is, $KL(p||q) \neq KL(q||p)$. This is not justified when comparing two samples from the same source.

4.2.1 Case example: interarrival times of GPRS packets

Figure 4.6 outlines a small-scale remote automation system, where a distant unit operates out of reach of wired networks. The unit itself may contain some stand-alone automation but it also sends maintenance data over GPRS network. The data sent to the management center could be, for example, update information of a graphical display, in which case the packets sent are small and frequent. Examples of a remote automation application include meteorological stations, harvesters and unmanned power plants.

Figure 4.7 helps understand the course of the case example. Data from a distant

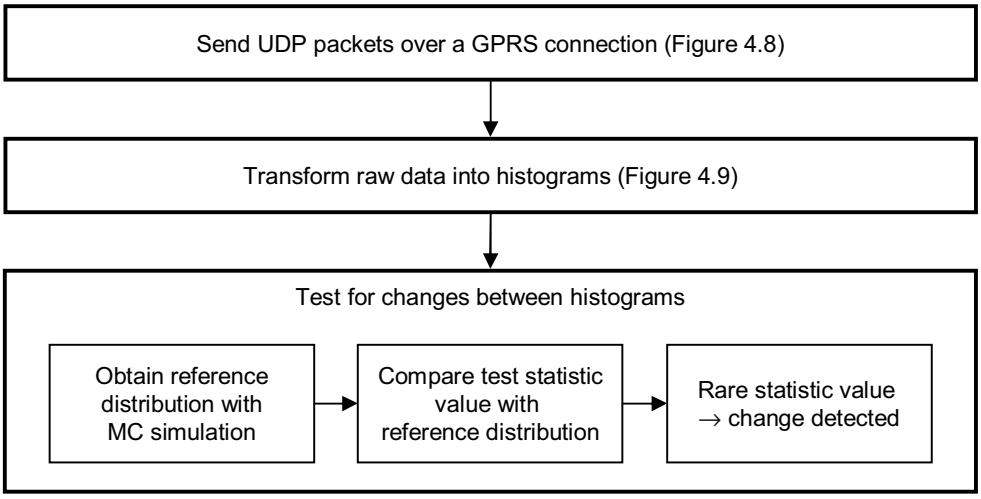

Figure 4.7: Course of the GPRS case example.

unit were simulated by sending small UDP packets with varying intervals over a true GPRS connection. The aim was to investigate whether a statistical test at the receiving end can detect a change in the sending rate with no other information from the sender. A change in the rate could indicate a device malfunction or a malicious attack in the sending unit [97], though also just a change in the amount of high-priority traffic overriding GPRS data. Deviations in the interarrival time can also be interpreted as jitter, which is an important Quality of Service parameter. In addition to multimedia applications, jitter is crucial to automation systems, especially if the connection is used for feedback control [37].

The simulation had four 200 packet phases (Figure 4.8). The time interval between packets at the sending host was normally distributed; Table 4.3 lists the parameters of the normal distribution for each phase.

Figure 4.8 reveals some strange phenomena that affected the study. When the deviation of the sending rate increased in phase 2, the arrival rate deviation seemed not to increase at all. This could be due to a buffer somewhere in the network. A typical behaviour of a buffer is to reduce jitter, which was exactly the result here. In phase 3 then, the interarrival times seemed to diverge into three distinct levels. One of the levels was remarkably lower than the sending rate, which can only

Table 4.3: Parameters of the normally distributed intermediate times between sent packets in the example case.

Phase	1	2	3	4
Mean	1.15	1.15	1.675	1.675
Standard deviation	0.05	0.1	0.05	0.1

be explained by a buffer from where several packets burst at once. This in turn inherently increased jitter, as is clear from the figure. The reason for the buffering phenomenon was most probably appearance of voice calls in the network, thanks to which GPRS best effort data traffic had to queue.

Because the purpose of this arrangement was to detect changes from histograms, the raw data of Figure 4.8 were not used as such but were transformed into histograms first. Now suppose that all information that was available from the data acquisition system was a 14-bin[1] histogram of the interarrival times. Figure 4.9 is an example histogram of the interarrival times of the fourth phase. Outliers, such as the lonely point at 3.2 seconds, widen the histogram and thus reduce its informativity.

Two tests were performed: one to test for a change between phases 1 and 2, and another one between phases 3 and 4. The change in the mean value between phases 2 and 3 was so distinct that it was not worth testing. The test procedure followed the theory presented in Chapter 2: The null hypothesis was that there was no change between the phases, that they came from the same distribution. To test this, the Kolmogorov-Smirnov statistic served as a measure of discrepancy.

Histograms like the one in Figure 4.9 were used for Monte Carlo simulation. A uniform random sample of 200 observations was transformed into a sample distributed similarly to the histogram using the inverse transform method introduced in Chapter 2. Then, the K-S statistic was calculated from 1000 such samples compared with the true data of phase 1. These 1000 statistics served as a reference

[1]Several studies suggest methods for selecting the number of bins, usually on the basis of the data. Since this example mimics an automation system that does not adapt to the characteristics of the data, no attention is paid to the bin count. The choice $\sqrt{n} \approx 14$ hardly affects the results.

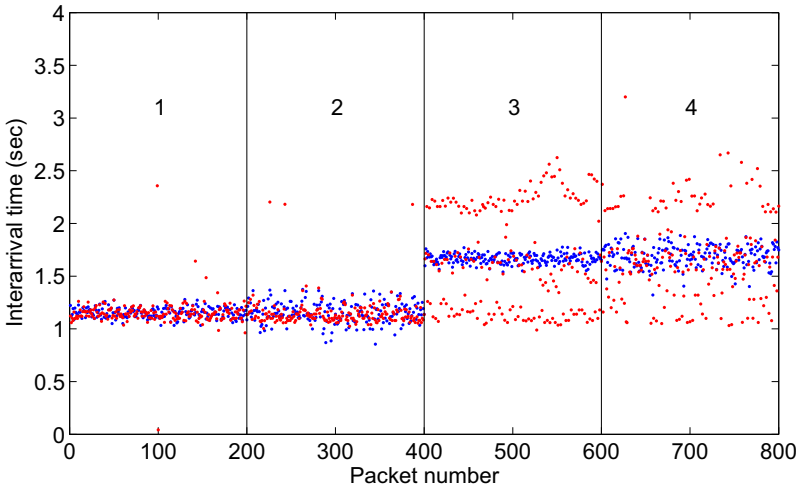

Figure 4.8: Inter-packet times of sent (blue) and received (red) GPRS data. The four phases had different characteristics of the sending rate (see Table 4.3). The raw measurements were not used in this form, they were first summarized as histograms (see Figure 4.9 for an example).

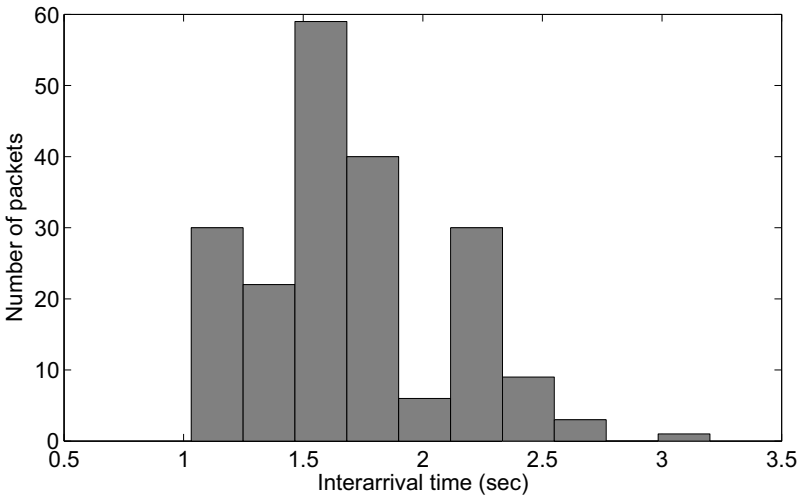

Figure 4.9: Histogram of the interarrival times of phase 4 (Figure 4.8).

distribution for the K-S statistic. If the null hypothesis “no change” was true, the statistic fitted in its reference distribution.

The K-S statistic distribution of phase 1 (Figure 4.8) was quite coarse. This was due to the discreteness of the variable: The cumulative distribution function of a discrete random variable only got values that were quotients of an integer and the sample size (200), which resulted in multiples of 0.005. The same applied to the K-S statistic, as it is calculated as a difference between two cumulative distribution functions.

Nevertheless, the cdf could be used for testing. The value of the K-S statistic between phases 1 and 2 was 0.015, and according to Figure 4.10 the P-value $\Pr\{x \leqslant 0.015\} \approx 0.50$ (the P-value could have been estimated numerically, but here looking at the figure will suffice). Thus, the statistic was not at all rare in its distribution and the null hypothesis could not be rejected. This confirmed the observation from Figure 4.8 that in spite of the increased jitter at sending, the receiver did not see the increase because a buffer en route dampened the jitter.

From the other phase, the result was different. Figure 4.11 shows the cdf of the K-S statistic simulated from phase 3. It is notably different from Figure 4.10 because the recorded interarrival times in phase 3 differed from those of phase 1. The value of the K-S statistic between phases 3 and 4 was 0.145, which got a P-value of 1 — in other words, all 1000 simulated statistics were smaller than 0.145. Even though the difference is visually not that clear (Figure 4.8), the statistical test rejected the hypothesis of similarity most certainly.

4.3 Conclusion

The data to be analyzed are often of discrete nature. This is particularly true for network traffic, where many phenomena are intrinsically discrete. Packets, bytes and connections are counted in integers, coarse measurements cause roundings, and, for example, network addresses and ports do not allow for numerical treatment at all. Furthermore, histogram is a common way to compress information.

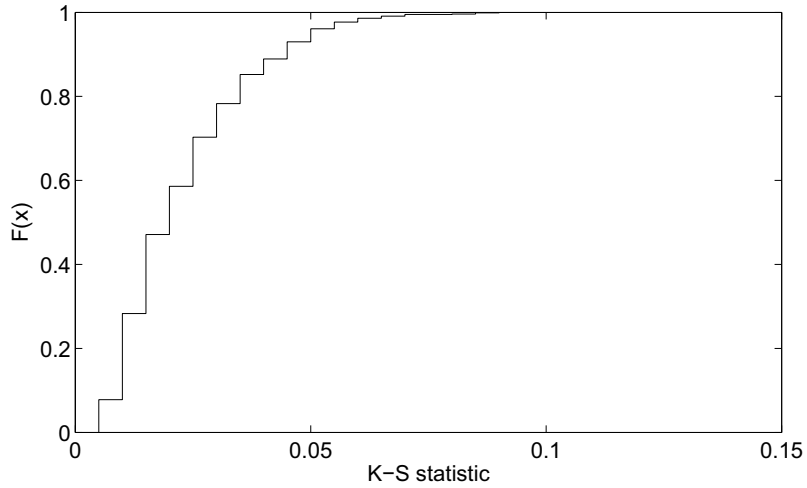

Figure 4.10: Cumulative distribution function of the Kolmogorov-Smirnov statistic simulated from phase 1 (see Figure 4.8).

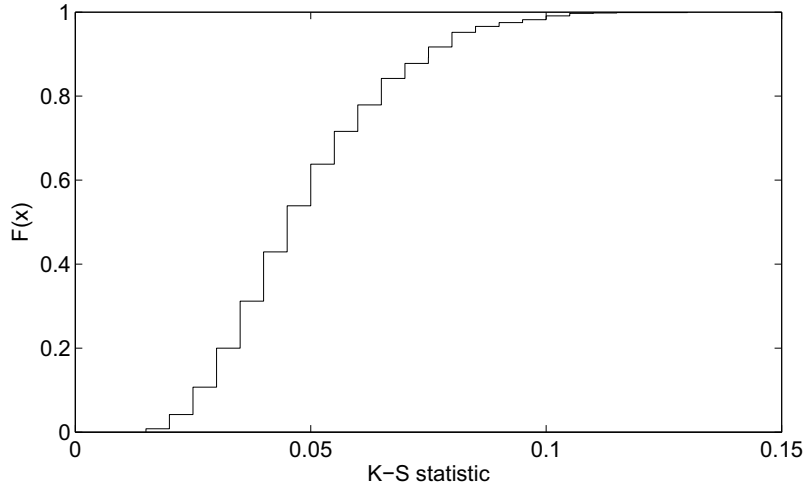

Figure 4.11: Cumulative distribution function of the Kolmogorov-Smirnov statistic simulated from phase 3 (see Figure 4.8).

This chapter used some commonly known statistical methods, but instead of continuous data, they were now applied to discrete data. The methods presented were able to detect changes in port histograms and GPRS data in histogram format. Some of the detected changes were somewhat obvious when verified by plotting illustrative figures. Once plotted, the changes would be easily observable to a human operator. Yet, no operator can analyze 50000 traces of network traffic, as there were in one of the cases. This is exactly were automated data analysis is needed: for exploring huge amounts of data collected by modern network management systems. The methods of this chapter automatically tell when a significant change occurs.

Many other methods might detect similar changes to the ones studied. However, the strength of statistical methods lies in their strong theoretical background. Furthermore, the method presented could be applied to another application with few modifications.

Chapter 5

Test statistic study

A goodness-of-fit test involves a test statistic, whose properties have to be known to some extent. The most popular tests use statistics that have analytically tractable distributions and thus exactly known properties. So far in this thesis, some tests have been cited as exact in some circumstances. In many true applications however, knowledge on the underlying process is limited and few assumptions can be made. When network traffic with its heavy-tailed distributions comes along, textbook versions of goodness-of-fit tests no longer apply; this problem was discussed in Chapter 3. This chapter addresses the problem of choosing a suitable statistical test for change detection of real network data. Parts of the results were published in [64] and will be published in [65].

Tests for heavy-tailed distributions have been presented in the literature [2, 103, 20, 23]; they usually resort to the Pareto distribution or its variants as an analytical approach. Still, exact or even asymptotic solutions are not always found, and Monte Carlo simulation must be used for finding the critical values. Simulation can handle more complex models with fewer assumptions than analytical solutions can [76].

A power study is a common means to investigate power, a salient property of a statistical test. This chapter reviews several test statistics and adds a few trivial ones, and conducts a power study to see which one to apply to the tricky distributions of network data.

Goodness-of-fit tests generally define the null hypothesis only, thus testing against "not null hypothesis". For determining the power, a more specific alternative hypothesis is necessary. A common arrangement is to test between two distribution families, such as exponential and Pareto distributions. This study takes a different approach: It filters protocols from real network data and uses them as an artificial traffic mixture. Both methods — using analytical distributions or real data — have their disadvantages and generalizations, but data collected from true networks are one of the guidelines of this thesis.

Three protocols were filtered from a NetFlow data file; flow sizes of these protocols are used in the study. First, HTTP flows serve as null hypothesis and Gnutella flows are added for finding out which tests detect the change. Then, the roles are swapped and it is seen if the results are different when the null hypothesis is Gnutella traffic. Finally, flows of the quite similar protocols Gnutella and Kazaa are used in another power study.

5.1 Introduction of test statistics

Eleven test statistics are employed in this study. Three of them (mean, median and maximum) are simple everyday statistics while the other eight were picked from publications where authors have proposed statistics for change detection of heavy-tailed distributions.

Smith [103] listed five test statistics collected from various sources. The original sources used the statistics for slightly different purposes; yet, Smith used them for detecting heavy-tailed distributions against a normal null distribution. Four of the statistics (K, U, V and W) contain the distance of each data point from the mean or median of the sample, which makes them sensitive to outliers. The fifth statistic (Z) uses the quartile means of the sample. Extremely large values typical to heavy-tailed distributions affect strongly all these statistics. Therefore, they probably work well in distinguishing heavy-tailedness from the normal distributions. Smith studied three sample sizes and various distributions from the Paretian family. He arrived at recommending different test statistics for different cases.

Smith's approach was fairly similar to this study, he used Monte Carlo simulation to estimate the cumulative distribution functions of the statistics under the null hypothesis. Even though the hypotheses in this study are different, the five statistics are adopted. They are the first five formulas in the list below. An obvious typing error in the formula of W in [103] has been corrected in Equation (5.4).

Brilhante [20] proposed a test statistic that she claimed to be resistant to outliers. She used the exponential distribution as a null hypothesis and tested against the generalized Pareto distribution. Brilhante's statistic proved to be inferior to some other alternatives in change detection, but it succeeded in resisting outliers and other contamination in the distribution. The resistance followed from using only quartiles and medians in the statistic, thus even up to 25 % of contamination does not affect the result. On the other hand, the robustness impairs the sensitivity of the test, particularly with heavy-tailed distributions where extremely large values are not necessarily contamination. Brilhante's statistic (C) as well as two other ones (A and B) she used are included in this study and listed below.

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4}{(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2)^2} \tag{5.1}$$

$$U = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}}{\frac{1}{n}\sum_{i=1}^{n}|X_i - X_{\text{med}}|} \tag{5.2}$$

$$V = \frac{X_{\max} - X_{\min}}{\frac{2}{n}\sum_{i=1}^{n}|X_i - X_{\text{med}}|} \tag{5.3}$$

$$W = \frac{X_{\max} - X_{\min}}{2\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}} \tag{5.4}$$

$$Z = \frac{\overline{X}_4 - \overline{X}_1}{\overline{X}_3 - \overline{X}_2} \tag{5.5}$$

$$A = \frac{X_{\max}}{X_{\text{med}}} \tag{5.6}$$

$$B = \frac{X_{\max} - X_{\text{med}}}{X_{\text{med}} - X_{\min}} \tag{5.7}$$

$$C = \frac{X_{(3n/4)} - X_{\text{med}}}{X_{\text{med}} - X_{(n/4)}}, \tag{5.8}$$

where

n is the sample size,

$\overline{X}$ is the sample mean,

$X_{(i)}$ is the ith order statistic,

$X_{\min} = X_{(1)}$ and $X_{\max} = X_{(n)}$ are the smallest and largest values of the sample, respectively,

X_{med} is the sample median, and

$\overline{X}_i$ is the mean of the ith quartile.

In addition to these somewhat complicated statistics, the simple statistics mean, median and maximum are included in the study.

5.2 Selection of data

This power study uses the real-world data described in Chapter 4. Three protocols; HTTP, Gnutella and Kazaa; were filtered from flows of a 5-minute data file. HTTP is the protocol for retrieving web pages, Gnutella and Kazaa are used by peer-to-peer (p2p) file sharing networks. The resulting flow counts are 27922 HTTP, 9404 Gnutella, and 8869 Kazaa flows. Though the number of HTTP flows is triple the number of the other two, even the smallest count is enough for analysis.

As was stated in the previous chapter, it is impossible to perfectly recognize all flows of a certain protocol. In this case, flows having ports 80, 443, 8000 or 8080 as either source or destination port were regarded as HTTP flows. The complementary cumulative distribution function (ccdf) of HTTP (Figure 5.1) shows a fairly linear slope on the double logarithmic scale, which indicates a heavy-tailed behaviour (cf. Figure 3.4).

Ports 6346 and 6347 signify Gnutella traffic [54], which contains both signalling traffic and data transfer. The ccdf of Gnutella reveals a bimodal distribution, where

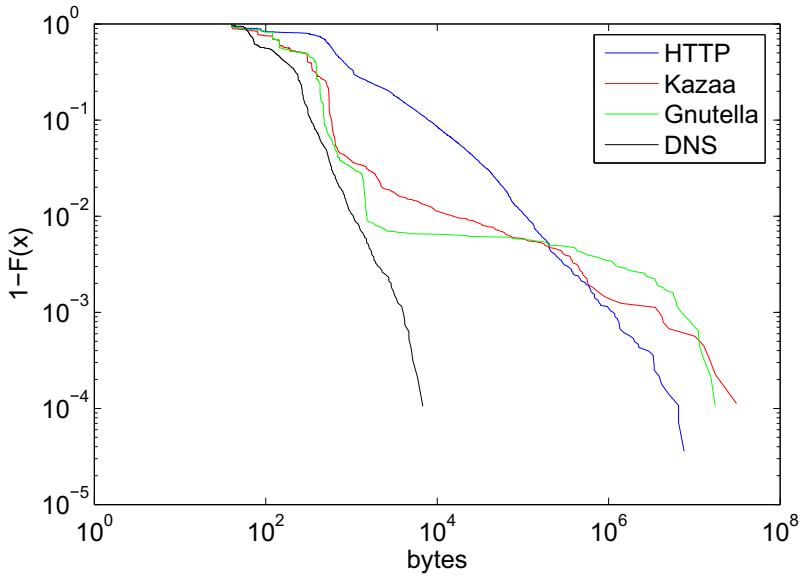

Figure 5.1: Complementary cumulative distribution function of HTTP, Kazaa, Gnutella, and DNS flow sizes.

a small fraction of flows (less than 1 %) carries large amounts of bytes compared to the others. This fraction is obviously data transfer, while the majority of flows belongs to signalling traffic. Ilie *et al.* [53] believed that the signalling traffic follow an exponential distribution; this could be true also from Figure 5.1. The transfer flows might exhibit heavy-tailedness, which however is impossible to judge from the figure. Part of the transfer probably travels through unknown ports, since the peers can agree on an arbitrary port after the desired file is found.

Kazaa uses port number 1214 [54]. As the two p2p-protocols are quite similar, what was said about Gnutella above applies largely to Kazaa too.

Figure 5.1 also presents the ccdf of DNS (Domain Name System) for comparison. DNS typically transfers only small flows that contain requests and responses of domain names and addresses [107], which can be seen from the curve. Detecting changes between DNS and the other three would be quite trivial because of the highly different distributions, thus DNS was left out of this study.

5.3 Test arrangement

Let the null hypothesis be that the traffic consist of pure HTTP flows. Stated like this, the hypothesis is unreal, but the target is to test for changes when one type of traffic is mixed with another. The alternative hypothesis is that the traffic contain a proportion of Gnutella among HTTP. For studying the power of the statistics under inspection, the proportion of mixed Gnutella traffic (θ) is first zero, that is, the null hypothesis is true, and then increases up to 1.

All of the statistics listed in Section 5.1 (excluding the three trivial ones) are somehow related to dispersion; they measure the dispersion of the random values. It is thus expected that the statistics decrease if the tail of the distribution gets lighter. However, it is hard to tell how the statistics behave when Gnutella traffic is mixed with HTTP traffic. This is why a two-tailed test must be carried out: A statistic deviating from its null distribution to either direction leads to rejection of the null hypothesis.

Recall from Chapter 2 that the power $P(\theta)$ of a test is the probability of not making a type II error, that is, the probability of rejecting a truly false null hypothesis. The power is usually a function of a parameter θ, in this case the proportion of Gnutella traffic in the true sample. The power of an ideal test would be zero when the null hypothesis is true and grow to one immediately when any Gnutella traffic is added. In practice however, the test is designed to reject even the true null hypothesis with probability α (significance), thus $P(\theta) = \alpha$ if the null hypothesis is true.

Finding out the power of a test for different values of θ includes two phases of Monte Carlo simulation. First, the empirical cumulative distribution of the test statistic T under the null hypothesis is simulated. Then, samples are simulated according to the alternative distribution. The proportion of rejections from these samples is the power. Finally, the procedure is repeated for each desired value of θ.

The following algorithm describes the course of the power study for one statistic. T denotes a generic test statistic in the algorithm; the algorithm is repeated for each statistic.

1. Let n be the sample size, k_0 the number of simulated null distribution samples, and k_1 the number of simulated test samples.
2. Draw k_0 samples of size n from HTTP flows. Calculate the test statistic T from each of the k_0 samples. Denote their cdf with F_T.
3. Let θ be the proportion of Gnutella flows in the sample.
4. Draw k_1 samples so that θn flows represent Gnutella and $(1-\theta)n$ represent HTTP in one sample. Calculate the test statistic T from each of the k_1 samples and determine their P-values using F_T.
5. Let $P(\theta) = \frac{1}{k_1}\sum_{i=1}^{k_1} 1_{P_i<\frac{1}{2}\alpha \mid P_i>1-\frac{1}{2}\alpha}$, where P_i is the P-value of sample i and 1_a gets the value 1 if the boolean expression a is true and 0 if it is false. $P(\theta)$ is now the power of the test. [1]
6. Repeat from 3 with several values of θ, $0 \leqslant \theta \leqslant 1$.

5.4 Results

The previous section described the test algorithm in detail. The algorithm was used not only in adding Gnutella to HTTP traffic but also vice versa. Change detection between Gnutella and Kazaa was attempted as well. This section illustrates the results.

5.4.1 HTTP as null hypothesis

Figures 5.2 and 5.3 show results of the power study with parameters $n = 100$, $k_0 = k_1 = 5000$ and $\alpha = 0.05$. The proportion θ of Gnutella flows was increased from 0 (null hypothesis) to 1 with 0.05 steps.

All statistics exhibit an increase in power when the proportion of Gnutella traffic increases. The quartile-based C behaves differently; its power decreases when

[1] This thesis follows the convention that both P-value and power are denoted by P. For the sake of clarity, power is written as $P(\theta)$.

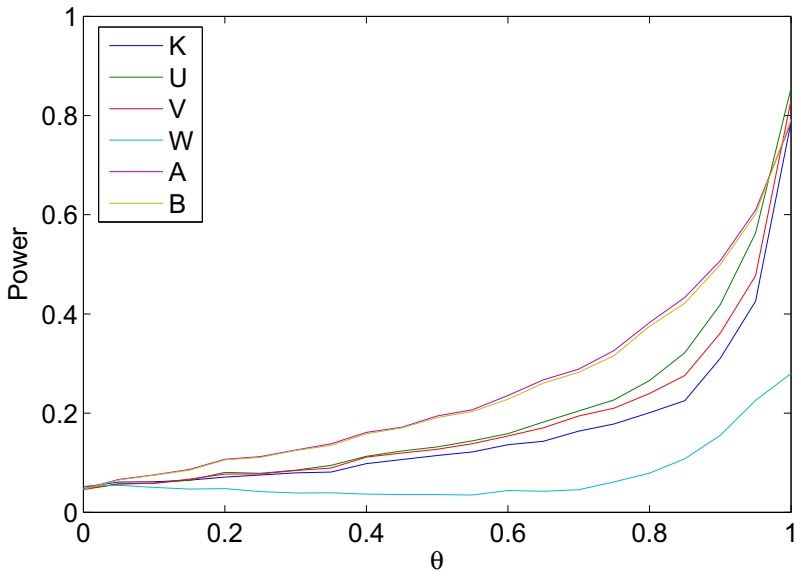

Figure 5.2: Power of test statistics K, U, V, W, A, and B when the null hypothesis is pure HTTP traffic ($\theta_0 = 0$) and Gnutella traffic is added.

θ approaches 1. However, the powers of all the eight complicated test statistics are more or less inadequate. The powers increase only moderately and approach unity only when almost all HTTP traffic is replaced with Gnutella traffic. Even the simple statistics mean and maximum do not lose to their more complicated counterparts. Furthermore, median seems to have the most optimal power curve of the eleven studied statistics. The power of the median reaches unity at $\theta = 0.4$, while no other statistic can show a power of 1 even at $\theta = 1$.

5.4.2 Gnutella as null hypothesis

In the above, the null hypothesis was HTTP traffic, and Gnutella flows were added as θ increased. It is possible to do the same study the other way around: Let the null hypothesis be that the traffic consist of pure Gnutella traffic ($\theta_0 = 1$), and HTTP flows are added. It is not at all self-evident that the results should look the same when the scenario is turned upside down, but still the results in Figure 5.4 surprise. The figure shows power curves of the same five statistics when all other

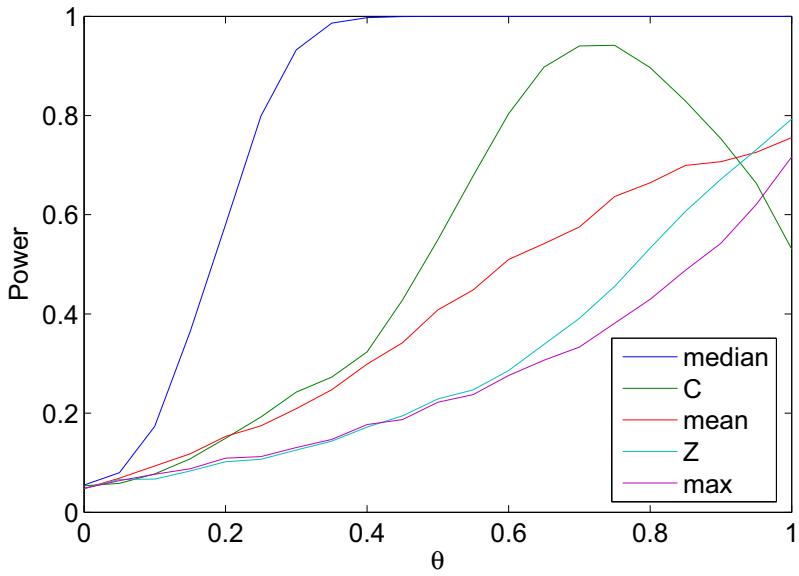

Figure 5.3: Power of test statistics Z, C, max, mean, and median when the null hypothesis is pure HTTP traffic ($\theta_0 = 0$) and Gnutella traffic is added.

parameters except the null hypothesis remained the same as in Figure 5.3.

Figure 5.4 reveals that only median can survive in detecting changes in both directions between Gnutella and HTTP traffic. The four other statistics in the figure have virtually no power at all, and the remaining six did not perform any better. Next, an explanation for these feeble results is sought.

Z is one of the test statistics that failed in the latter power study. Let Z_g and Z_h denote Z calculated from Gnutella (null hypothesis) and HTTP, respectively. Figure 5.5 depicts the cumulative distribution functions of Z_g and Z_h, both calculated from 5000 samples of 100 flows. Even the test statistics appear to show some heavy-tailed behaviour — the blue line is hardly visible in the upper left corner — but nevertheless, the curves show that Z_g has a much wider range than Z_h. Thus, when the null hypothesis rests on Gnutella, the high variability of Z_g drowns the variability in the alternate Z_h, and the null hypothesis hardly ever gets rejected.

The high variability of Z_g arises from the shape of the distribution of Gnutella

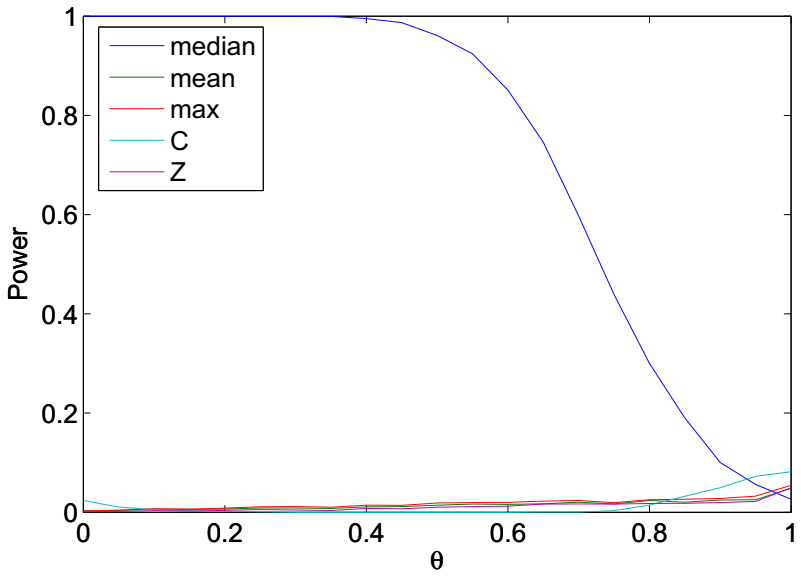

Figure 5.4: Power of test statistics Z, C, max, mean, and median when the null hypothesis is pure Gnutella traffic ($\theta_0 = 1$) and HTTP traffic is added.

flow sizes (Figure 5.1). The distribution contains a small fraction (less than one per cent) of very large flows that form a distinct lump in the ccdf curve. Upon sampling of 100 flows from the total population of 9404 flows, sometimes an extremely large flow gets into the sample and sometimes not. The largest value of the sample strongly affects $\overline{X}_4$, the mean of the highest quartile (see Section 5.1), and fluctuation follows. A similar mechanism most likely spoils also the chances of the other test statistics, since most of them are somehow prone to extreme values. The median is an exception, it is not affected even by large changes in the upper tail of the distribution, and this property makes it suitable for detecting changes in tricky distributions.

5.4.3 Change detection between Gnutella and Kazaa

Gnutella and Kazaa are highly similar in nature; both of them are p2p-protocols and transfer same content types. Distinctions may exist in signalling traffic, yet Figure 5.1 shows little difference in the ccdf's of these two protocols. Thus, it is

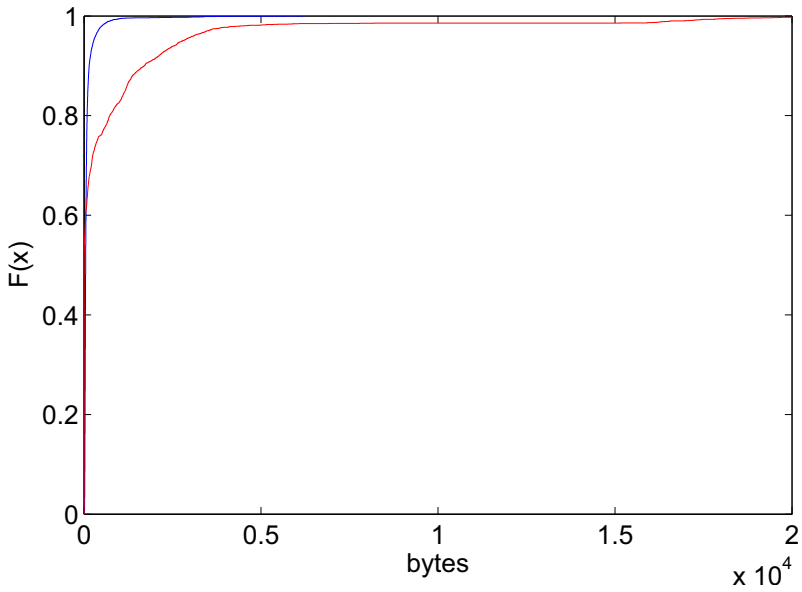

Figure 5.5: Cumulative distribution function of test statistic Z with HTTP (blue) and Gnutella (red) traffic.

presumable that no statistical test can adequately detect changes between Gnutella and Kazaa.

Figure 5.6 shows some results from the power study between Gnutella and Kazaa. The curves in the figure belong to three statistics that performed best when Gnutella traffic served as null hypothesis, although their performance is not even satisfactory. All other statistics showed no power at all. In the reversed study, that is, when Kazaa was used as null hypothesis, all eleven statistics failed. Hereby it is easy to conclude that this kind of a statistical test cannot detect changes between Gnutella and Kazaa.

5.5 Conclusion

A statistical test often equates with the test statistic used. Consequently, the choice of the test statistic plays a fundamental role in designing the test. This chapter re-

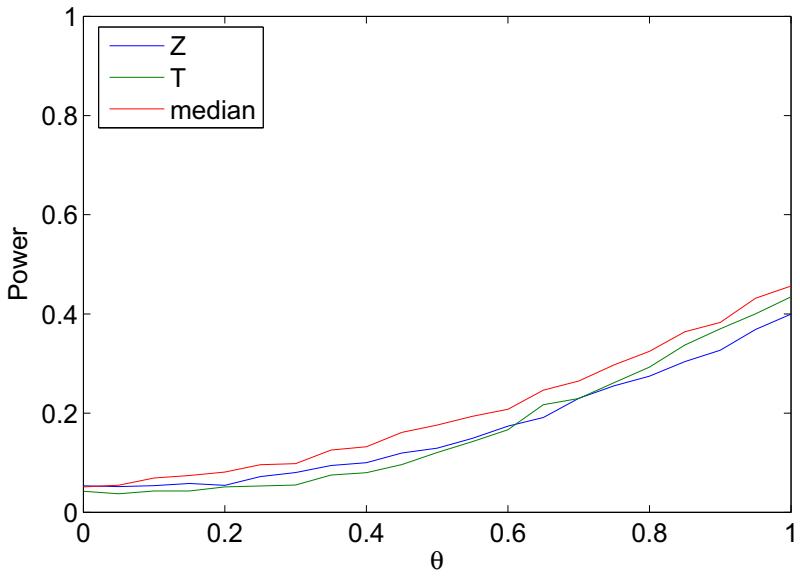

Figure 5.6: Power of test statistics Z, T, and median when the null hypothesis is pure Gnutella traffic and Kazaa traffic is added.

viewed several statistics connected with heavy-tailed distributions in the literature and compared them by means of a power study. Distributional assumptions and analytical solutions had to be discarded because the study used real traffic data with inconvenient statistical properties — Monte Carlo simulation was used instead. Detecting changes in an artificial mixture of real data turned out to be hard to most of the statistics studied.

The test arrangement used two traffic types collected from real traffic traces. With these two types, an artificial traffic mixture was created and changes in the mixture were detected. The results showed that most test statistics are too sensitive to the extremely large values typical to heavy-tailed distributions. Only median managed to distinguish both HTTP from Gnutella and Gnutella from HTTP. A further study investigated change detection between the very similar protocols Gnutella and Kazaa; this detection turned out impossible.

Certainly, the somewhat false arrangement with only a few protocols does not tell the whole truth. Nevertheless, the protocols HTTP and Gnutella were chosen from

the most popular ones in the Internet, yet different enough by nature to be distinguished. In traditional data analysis, outliers are considered faulty measurements and removed, but in network traffic data they belong to the process. If a statistical test is prone to extreme values, it should be replaced. Though the median in its simplicity worked well, more sophisticated test statistics based on median could be developed.

The choice of the test statistic is not the only crucial one. The next chapter examines the importance of sample size.

Chapter 6

Sample size study

Earlier in this thesis, several samples were drawn and tested with various methods. Sample sizes were chosen on intuitive grounds with fairly little consideration. In this chapter, it is time to pay more attention to the sample size, one of the most important parameters — though not always selectable — in statistical testing. Parts of the results will be published in [65].

Virtually all textbooks and other publications that discuss sample sizes worry about the sample being *too small*. This is the natural point of view when observations are slow, cumbersome or expensive to make. Medical surveys, for example, sometimes involve years of follow-up, or geographical exploration may require drilling deep holes into the soil. Too small a sample would impair the statistical significance of the results and possibly even ruin the whole research. On the other hand, making excessive observations just to be on the safe side would be waste of resources. Under these conditions, taking a sample big enough but no more is of utmost importance.

Everything turns upside down when there are more than enough data available. In a typical network traffic data analysis case, including 1000 observations instead of 100 does not notably increase costs. This is particularly true with network management systems, where huge amounts of measurements are stored anyway. Cheap storage space has made collecting data perhaps even too easy; finding relevant information from the data becomes a problem. Collecting samples *not too*

big helps in distinguishing true changes from negligible noise.

Barnes [12] discussed *practical significance* in contrast to statistical significance. He claimed that a statistically significant deviation from the null hypothesis can always be found if the sample is large enough. Practical significance means designing the test and sample size so that negligible deviations will not reject the null hypothesis. Barnes compared too big samples with a botanist who studies colour variations of a flower petal with a microscope when a magnifying glass would be more appropriate. In the telecommunication world, a similar example might be measuring hourly traffic volumes with a packet-level analyzer instead of a trend display.

Important parameters of a statistical test include significance α, effect size d and power $P(\theta)$. The significance level has received plenty of attention in the previous chapters, but the other two parameters are closely related to the subject of this chapter. The effect size "reflects an alternative of scientific interest" [69], that is, deviations from null hypothesis smaller than the desired effect size are not of practical significance. The word "effect" has its origin in medical sciences, where a treatment is expected to have an effect.

The target value of power is the desired probability of a truly false null hypothesis getting rejected at effect size d. For example, a statistical test might test whether the average delay of a connection exceeds a given threshold. The effect size could be 100 ms (above the threshold) and the desired power 80 %, which is a common value for power [69]. Now the sample size should be determined so that a sample where the average delay exceeds the threshold with 100 ms would lead to rejection of the null hypothesis with a probability of 80 %. The notation $P(\theta)$ underlines the fact that power always depends on some parameter of the true underlying distribution, in this case the average delay.

Guidelines for determining sample sizes exist, see for example [102]. These guidelines always rest on assumptions, either distributional assumptions or parameter estimates. This chapter avoids such assumptions and uses past data for simulation. Of course, relying on historical data is as much an assumption as any other; the results of the simulation study in this chapter are valid only for a case like the one simulated. However, the results will provide valuable information on the connection between sample size and power.

The chapter contains two parts. First, an example illustrates the meaning of sample size to goodness-of-fit testing. Then, the actual power study with different sample sizes is presented. Some concluding remarks and guidelines are provided at the end.

6.1 Example: information content of a large sample

Consider a standard normal distribution, that is, $\mu = 1$ and $\sigma = 1$. Figure 6.1 shows a sample of 100 observations, where 90 observations come from a standard normal distribution and the other 10 from a normal distribution with $\mu = 0.9$, $\sigma = 1$. The "disturbing" 10 observations are marked with a circle, and they hardly stand out from the rest of the sample.

Indeed, the mixed sample of 100 points cannot be distinguished from a pure standard normal distribution by means of a goodness-of-fit test (Table 6.1). The Anderson-Darling statistic calculated from the sample remains well below the appropriate critical value ($\alpha = 0.05$), when testing for normality. When the sample size is increased, the inreasing information content makes the A^2 statistic grow rapidly. With a sample of $n = 100\,000$, there is no doubt about the impurity of the hypothesized normal distribution. If the large sample were plotted, it would look roughly the same as Figure 6.1, only on a different scale. Yet the large sample can be almost certainly identified as non-normal — no matter if the difference is of practical significance or not.

6.2 Test arrangement

Now let us move further to studying the effect of sample size with real traffic data. The test arrangement here is most similar to that of the previous chapter; HTTP and Gnutella serve as examples of diverse network protocols. As a result of the previous chapter, median is first used as the test statistic and only the sample size is varied.

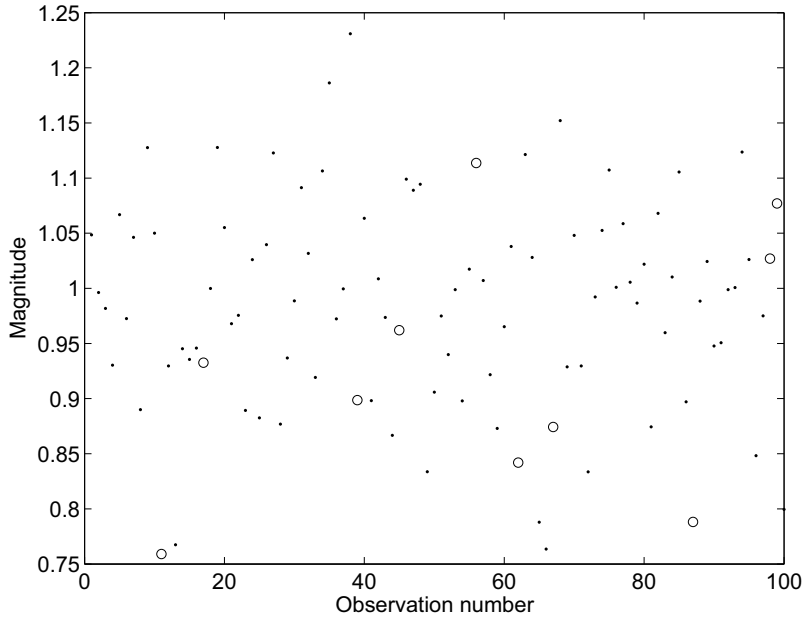

Figure 6.1: 90 observations from a normal distribution with mean 1.0 (dots) and 10 observations from a normal distribution with mean 0.9 (circles). Both distributions have unit variance.

Figure 5.1 of the previous chapter presented complementary cdf's of Gnutella and HTTP. The figure showed that Gnutella traffic contains a small fraction of large flows, while most Gnutella flows are smaller than HTTP flows. Since the fraction of large flows is less than one per cent of all Gnutella flows, it probably does not affect the median. Thus the median of HTTP flow sizes is larger than that of Gnutella flows, which suggests that the median as a test statistic decrease when the proportion of Gnutella increase. This would recommend a one-tailed test, but in conformity with the previous chapter, a two-tailed test is used here too.

Again, let θ denote the proportion of Gnutella flows in the sample. The null hypothesis is pure HTTP ($\theta = 0$), and θ increases gradually to build up the power curve. Two rounds of Monte Carlo simulation are needed for each sample size studied; one for creating the null distribution and one for calculating the power. The following algorithm describes the course of the sample size study in detail.

Table 6.1: Detection of an impure normal distribution with Anderson-Darling test and samples of different sizes.

Sample size	A^2	critical value
10^2	0.230	0.752
10^3	0.333	0.752
10^4	0.775	0.752
10^5	4.188	0.752

1. Let k_0 be the number of simulated null distribution samples and k_1 the number of simulated test samples.
2. Let n be the sample size.
3. Draw k_0 samples of size n from HTTP flows. Calculate the median from each of the k_0 samples. Denote the cdf of the medians with F_{med}.
4. Let θ be the proportion of Gnutella flows in the sample.
5. Draw k_1 samples so that θn flows represent Gnutella and $(1-\theta)n$ represent HTTP in one sample. Calculate the median from each of the k_1 samples and determine their P-values using F_{med}.
6. Let $P(\theta) = \frac{1}{k_1}\sum_{i=1}^{k_1} 1_{P_i<\frac{1}{2}\alpha \mid P_i>1-\frac{1}{2}\alpha}$, where P_i is the P-value of sample i, and 1_a gets the value 1 if the boolean expression a is true and 0 if it is false. $P(\theta)$ is now the power of the test.
7. Repeat from 4 with several values of θ, $0 \leqslant \theta \leqslant 1$.
8. Repeat from 2 with several values of n.

6.3 Results

Figure 6.2 shows the results for $k_0 = k_1 = 5000$, $\alpha = 0.05$, and several values of n. The ruggedness of the gentlest curve is due to the small sample: When the sample

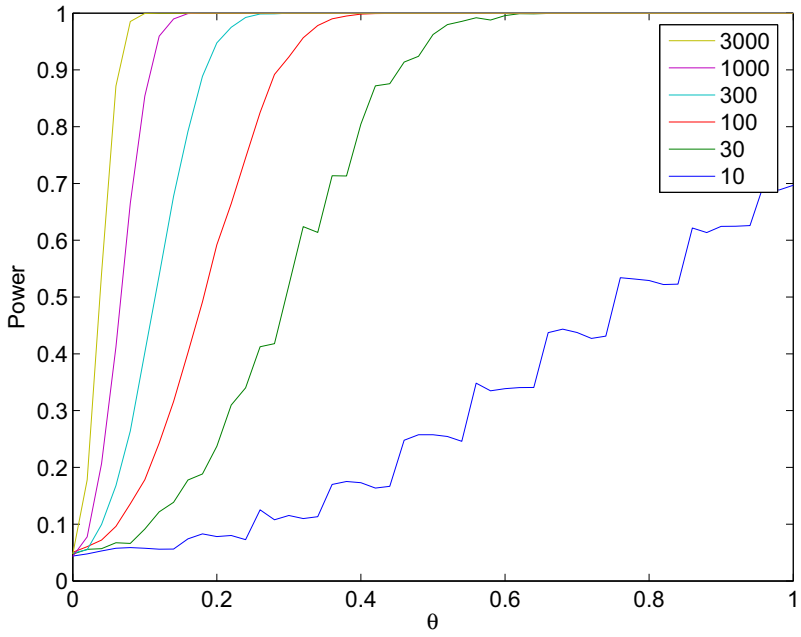

Figure 6.2: Power curves of test statistic median with various sample sizes.

contains only 10 observations, $\theta = 0.38$ and $\theta = 0.42$, for example, do not differ and the resulting power may vary randomly with the simulations.

Power curves like the ones in Figure 6.2 help the data analyst in selecting the correct parameters. The largest sample ($n = 3000$) provides a unit power as early as $\theta = 0.1$, so only a 10 % portion of Gnutella traffic among HTTP makes the test positively reject the null hypothesis of pure HTTP traffic. This sounds quite a rude judgment.

The power of $n = 100$ shows a power of 0.8 at approximately $\theta = 0.3$, which means that with a sample of one hundred observations, the median test can distinguish 80 % of samples with 30 % Gnutella traffic from pure HTTP. Furthermore, if the percentage of Gnutella grows to 40, the sample is almost certainly big enough to reject the hypothesis of HTTP traffic. One hundred observations appears to be a good compromise for the sample size.

Examining two additional test statistics from Chapter 5 gives further information

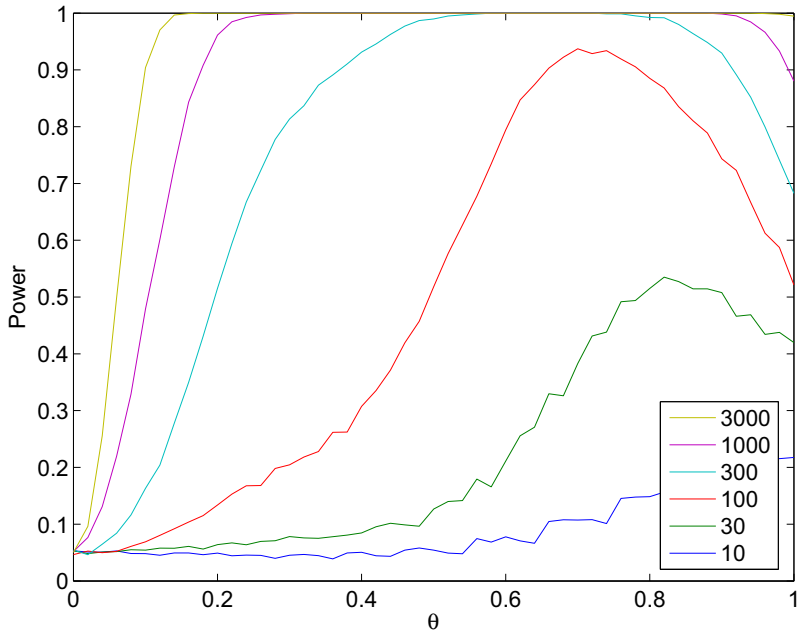

Figure 6.3: Power curves of test statistic C with various sample sizes.

on the effect of sample size. Figures 6.3 and 6.4 present the sample size study results of test statistics C and Z, respectively.

With large sample sizes, C seems to perform well. The power rises to unity quite rapidly. Still, the decreasing power at θ values close to 1 remains even with big samples. This does not make C an attractive choice for change detection.

The test statistic Z does not benefit very much from increasing the sample. Only the sample size $n = 3000$ yields a power that tells of a tolerable change detection ability, but even that cannot compete with the other two.

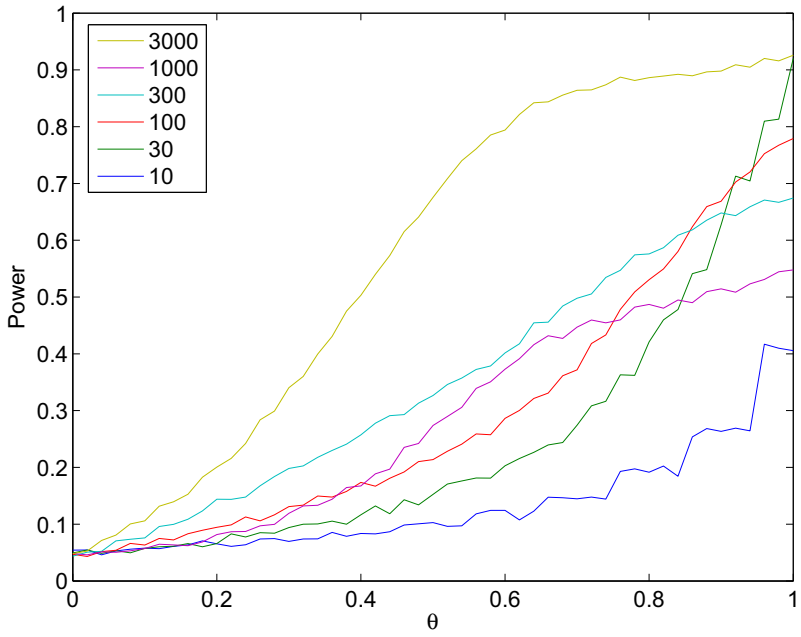

Figure 6.4: Power curves of test statistic Z with various sample sizes.

6.4 Conclusion

Sample size determination is always a compromise. It involves parameters that depend on each other — such as significance, power and effect size — and do not always have straightforward interpretations [10]. Thus any results should be considered only suggestive.

This chapter approached the sample size problem without strict calculations, but the outcome remains the same. The final decision is up to the analyst and depends on the choice of other parameters. Compared to other sample size studies, this one was based on the premise that sampling is neither slow nor expensive. In contrast to many other environments where excessive sampling is waste of resources, a network management system inherently collects huge amounts of data. Thus the viewpoint presented in this chapter is valuable when choosing the appropriate sample size.

6.4.1 Guidelines for change detection with goodness-of-fit tests

Chapters 5 and 6 discussed some important choices concerning change detection using goodness-of-fit tests. The following overview sums up some of the findings from a very practical viewpoint. The choices made by a data analyst always depend on the data and application, but some general guidelines can be found from the results of this and the previous chapter.

- Real network traffic data is often hard to analyze with mathematically analytical methods. Instead, Monte Carlo simulation offers means to find numerical answers with computational power.
- Median worked well as a test statistic. Additionally, another test statistic that was based on median gave tolerable results. Heavy-tailed distributions entail extremely large values that can be handled with median.
- The effect size has to be decided case-specifically. In these studies the effect size was not strictly defined, but in the background there was an estimate of effect size $\theta \approx 0.3$ and power $P(\theta) \approx 80$ %. Put in words, changing the proportion of a certain traffic type by 30 % should be detected with a power of 80 %. The former choice is very case-specific, while the latter is a commonly used power target.
- A big sample allows detecting small changes. But as processes always contain also changes that should be neglected, sample size should be carefully chosen and possibly even iterated. The results of this chapter showed that a sample of less than 100 observations is hardly sufficient when observations are virtually free of costs. $n = 100$ worked well in general, but even $n = 1000$ is not necessarily too much.

Chapter 7

Heavy-tailed distributions in traffic prediction

Together with its side-effects, heavy-tailed distributions have been considered a nuisance in network management. As discussed in Chapter 1, heavy-tails relate to sudden extremely large amounts of traffic or prolonged bursts, which pose problems to designing network protocols, buffer sizes and topology.

However, Park and Willinger [84] demonstrate an interesting benefit of heavy-tailed distributions: They lend themselves to predicting better than the more familiar distributions like the exponential. Many authors [60, 88, 101] have referred to the supposed predictability, but no one has actually used it. Instead, many other approaches to predicting heavy-tailed or self-similar traffic do exist.

Autoregressive (AR) and autoregressive moving average (ARMA) models are among the simplest dynamic time series models. Several studies that predict network traffic with AR-models have been published. Shah *et al.* [99] predicted simulated network traffic with various linear and nonlinear AR(MA) models. They used the ns-2 network simulator with two network topologies and Pareto distributions for simulating flow sizes. Their main result was somewhat surprising: The dynamic predictors did not show much improvement over such simple predictors as the last observation or the mean of the time series.

Also Baryshnikov *et al.* [13] trusted in the strength of simple predictors and used linear extrapolation. They used interesting traces, such as web server requests from the winter olympics and football world cup, and predicted workload peaks that they called hotspots. According to the results, the designed method based on linear extrapolation worked well and proved robust in predicting server overloads.

Xiaohu *et al.* [119] started from the fact that variance is infinite for heavy-tailed distributions with tail indexes smaller than 2. They therefore substituted covariation for covariance and derived a method for identifying the parameters of a linear AR-model. According to the authors, the numerical experiments with traces from [67] showed that the method can predict changes in the burstiness of the traffic. However, the article omitted comparisons to other possible prediction methods. Judging by the figures in [119], similar results might have been achieved by much simpler methods.

Hinich and Molyneux [50] proclaimed network traffic nonlinear and doubted whether current techniques are sufficient for predicting the traffic. By "current" techniques they presumably meant linear AR-models, which cannot capture the nonlinear behaviour. Simple models are certainly inadequate for predicting self-similar traffic, but far more sophisticated methods can be found. Hinich and Molyneux also questioned the existence of long-range dependence, which is contrary to many other results presented in the literature [110, 3, 58].

Andreolini *et al.* [9] designed a whole framework around change detection and prediction of web server load. The framework contained two phases: firstly, a filtered representation of the measurements, referred to as the load tracker, and secondly, a separate block for change detection or prediction. The authors preferred using load trackers instead of direct measurements to help reduce high variability of the measurements and computational complexity. The predictor part of the framework consisted of a simple linear prediction that was not a very intelligent one. Yet, the power of the two-phase framework is the possibility to have the predictor replaced by any other method.

Along with the AR and other elementary methods, more sophisticated methods of network traffic prediction include generalized autoregressive conditional heteroskedasticity (GARCH) and wavelets. In their recent paper [7], Anand *et al.* used a GARCH model for predicting the bit rate of a real-world traffic trace.

Thanks to its variable variance, the GARCH model could capture the bursty nature of the traffic better than some other alternatives. Wang and Shan [116] combined wavelets and recursive least squares (RLS) to decompose long-range dependent traffic into short-range dependent traffic. In spite of the authors' confidence, the results are presented quite unclearly.

The above-cited references mostly referred to the complexity of heavy-tailed or self-similar traffic and tried to predict it with more or less classical methods. Conversely, He *et al.* [47] turned long-range dependence (LRD) into an advantage. They created a simple predictor that copies the structure of long-range dependent autocorrelation function and thus gives weight to samples far in the history. They also designed a TCP congestion control scheme that uses the prediction method.

Östring and Sirisena [81] utilized the LRD structure for prediction in a way slightly similar to [47]. But in contrast, they concluded that the long history is *not* the main reason behind predictability. Rather, they demonstrated that primarily the short-term correlations dominate the performance of the predictor.

Unlike all previous studies, this chapter concentrates on the presumed predictability as an inherent feature of heavy-tailed distributions. The target is to use the heavy-tailedness of flow durations for predicting traffic volume. The Pareto distribution is used as an example in the calculus as well as in the simulations, since real-world data from very specific applications are hard to obtain.

The organization of the chapter needs some introduction (Figure 7.1). After some basic definitions, the probability that a flow remains open is derived. As the probability is a function of the flow age, the distribution of the flow age is deduced. With these tools, the distribution of the probability is found and a model of open flows built. Finally, the probability is studied as a function of the prediction horizon and parameters of the distribution and thus viewed as predictability of the flows.

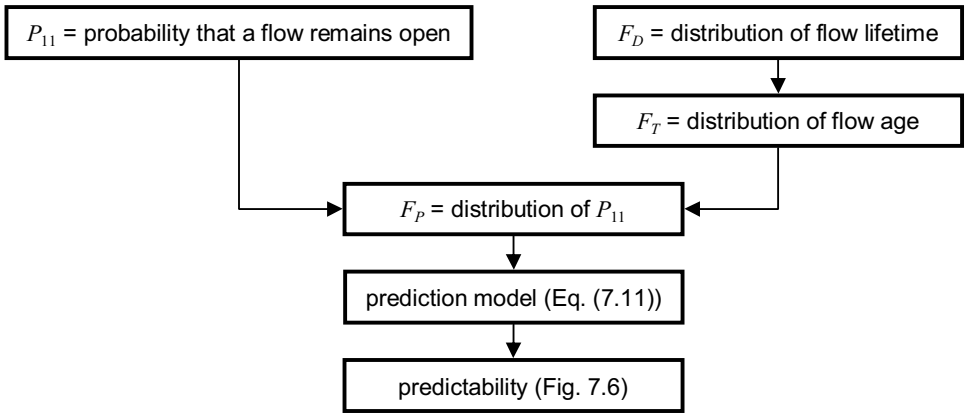

Figure 7.1: Organization of this chapter.

7.1 Heavy-tailed distributions

As was stated in Chapter 2, a random variable X has a heavy-tailed distribution if its cumulative distribution function (cdf) $F(x)$ has a tail exhibiting a power-law behaviour:

$$F(x) = \Pr\{X \leqslant x\} \sim 1 - kx^{-a} \tag{7.1}$$

In the above, $\Pr\{\cdot\}$ denotes probability, k is a constant coefficient and a is a constant called tail index or shape parameter. The term heavy-tailed describes the behaviour of the distribution at extremely large values of x: a heavy-tailed distribution decays slower than, for example, normal or exponential distributions, which can be termed light-tailed. [84] Sometimes the tail is turned downward by specifying the complementary cdf $\overline{F}(x) = 1 - F(x)$.

The Pareto distribution is a widely used heavy-tailed distribution. Its cdf and pdf were presented in Chapter 2. The mean or expected value of the Pareto is

$$E\{x\} = (ab)/(a-1) \tag{7.2}$$

if $a > 1$, where a is the shape parameter and b the location parameter ($x \geqslant b$). If $a \leqslant 1$, the expected value is infinite. The variance is finite only if $a > 2$, which is not common in practical applications. Hence, for the most typical tail index values $1 < a < 2$, Pareto-distributed variables (and, as a matter of fact, all heavy-tailed variables) have finite means but infinite variances. [84]

The subsequent analysis will use the Pareto distribution because of its simplicity. Additionally, the exponential distribution serves as a counterexample. The cdf of the exponential was introduced in Chapter 2.

7.2 Predicting flow durations

Suppose that the duration of a network connection follows a heavy-tailed distribution. Then, the longer the connection has been active, the more likely it is to remain active. Whereas Park and Willinger [84] only show the result in the discrete-time case, the following deduction extends the principle into continuous time.

Let the continuous random variable D represent the duration of a network connection, a flow (for example, the download time of a web page). The probability that the flow remains active after time period h, given that it has been active for time T, is

$$P_{11}(T,h) = \Pr\{D \geqslant T+h \mid D \geqslant T\} = \frac{\Pr\{D \geqslant T+h\}}{\Pr\{D \geqslant T\}}, \tag{7.3}$$

where the latter equality follows from the definition of conditional probability [78], since obviously $(D \geqslant T+h) \cap (D \geqslant T) = (D \geqslant T+h)$.

Because the probability of a single value of a continuous random variable is zero, $\Pr\{D \geqslant T\} = \Pr\{D > T\} = 1 - \Pr\{D \leqslant T\}$. Thus (7.3) can be written as

$$P_{11}(T,h) = \frac{1 - F_D(T+h)}{1 - F_D(T)}, \tag{7.4}$$

where $F_D(x)$ is the cdf of D.

Suppose now that the flow duration D is exponentially distributed. Substituting the exponential cumulative distribution function into (7.4) yields

$$P_{11}(T,h) = \frac{1 - 1 + \mathrm{e}^{-\frac{T+h}{\mu}}}{1 - 1 + \mathrm{e}^{-\frac{T}{\mu}}} = \mathrm{e}^{-\frac{h}{\mu}}, \tag{7.5}$$

where μ is the (only) parameter of the exponential distribution. The probability P_{11} is thus independent of T, the history of the flow being active. This result is closely related to the fact that the exponential is the only memoryless distribution [82]. If flow durations follow the exponential distribution, their finish rate is constant anytime.

A heavy-tailed distribution in turn behaves totally different from the above. Take the Pareto as an example and substitute its cdf into (7.4):

$$P_{11}(T,h) = \frac{1 - 1 + \left(\frac{b}{T+h}\right)^a}{1 - 1 + \left(\frac{b}{T}\right)^a} = \left(\frac{T}{T+h}\right)^a \tag{7.6}$$

The probability that the connection remains active thus depends on T, the time it has been active. Furthermore, the probability tends to unity as T tends to infinity. Hence, the longer a flow has been active, the more likely it is to remain active after h time units. This is a tempting result since it suggests that heavy-tailedness of the distribution could be exploited in predicting durations. This possibility is now further investigated.

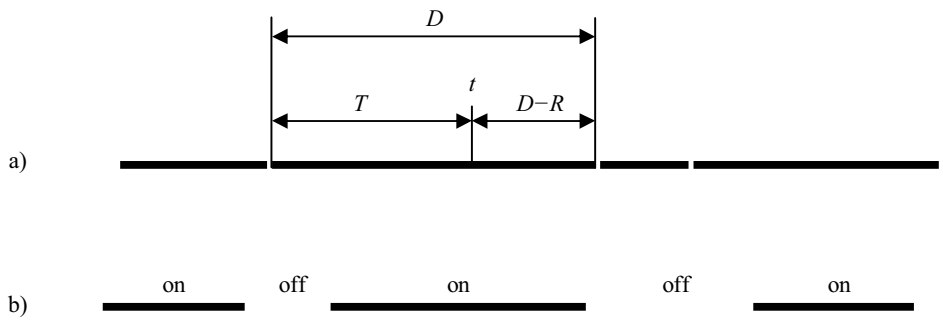

Figure 7.2: a) Renewal process. $D =$ flow lifetime, $T =$ flow age, $D - R =$ residual lifetime, $t =$ random time instant. b) Alternating renewal process.

7.2.1 Distribution of flow age

The distribution of the random variable T in (7.6) is not equal to the distribution of D. To clarify the distinction, T is called the *age* of a flow and D its *lifetime*. If a flow is active at a random time instant t, then $T \leqslant D$ but D is unknown.

Another assumption is needed for getting a grip of the distribution of the flow age T. A *renewal process* [11] is a process where a new flow begins immediately after another (Figure 7.2 a). For a renewal process, the cdf of T at the limit $t \to \infty$ can be expressed as

$$F_T(x) = 1 - \frac{1}{\mu} \int_x^\infty 1 - F_D(u)\,\mathrm{d}u, \tag{7.7}$$

where F_D is the cdf of the flow lifetime D and $\mu = \mathrm{E}\{D\}$ the expected value of it. [94] The limit $t \to \infty$ is of interest since in a real-world network the start time of the process is not known.

The remaining lifetime of a flow at a random time instant, $D - T$, is called the *residual lifetime* of the flow. Interestingly, flow age and residual lifetime are

equally distributed [94]. Anyway, here the flow age distribution F_T will be used to model the probability P_{11} in (7.6).

Substituting the exponential cdf $F_D(x) = 1 - \mathrm{e}^{-x/\mu}$ into (7.7) yields $F_T \equiv F_D$, which is not surprising. The residual lifetime is at any time instant distributed equally to the whole lifetime, in other words the memoryless exponential distribution does not remember how long the flow has been active.

For the Pareto distribution $F_D(x) = 1 - \left(\frac{b}{x}\right)^a$, $x \geqslant b$ the distribution function F_T has to be defined piecewise for x values less or greater than b, because $F_D(x) = 0$ when $x < b$. Given $1 < a < 2$, Equation (7.7) becomes

$$F_T(x) = \begin{cases} 1 - \frac{a-1}{ab}\left(\int_x^b 1 \,\mathrm{d}u + \int_b^\infty \left(\frac{b}{u}\right)^a \mathrm{d}u\right) = \frac{a-1}{ab}\, x, & \text{when } 0 < x < b \\ 1 - \frac{a-1}{ab}\int_x^\infty \left(\frac{b}{u}\right)^a \mathrm{d}u = 1 - \frac{1}{a}\left(\frac{b}{x}\right)^{a-1}, & \text{when } x \geqslant b. \end{cases} \tag{7.8}$$

Equation (7.8) reveals some properties of flow ages. If the flow lifetimes follow a heavy-tailed distribution with a tail index $1 < a < 2$, the tail index $a-1$ of the flow age distribution falls into the interval $(0,1)$. Consequently, the flow age distribution has an infinite mean (see Section 7.1). This might be the first hint suggesting that heavy-tailed distributions may not be a prominent tool in this kind of predicting.

For subsequent calculation of expected values, also the probability density function of T, the derivative of (7.8), is needed:

$$f_T(x) = \begin{cases} \frac{a-1}{ab}, & \text{when } 0 < x < b \\ -\frac{(-a+1)\ b^{a-1}}{a}\, x^{-a}, & \text{when } x \geqslant b \end{cases} \tag{7.9}$$

Figures 7.3 and 7.4 show the distribution and density functions of the flow age T, respectively. Parameter values used in the figures are $a = 1.2$ and $b = 1$. The probability density function has a logarithmic scale to improve clarity. The constant

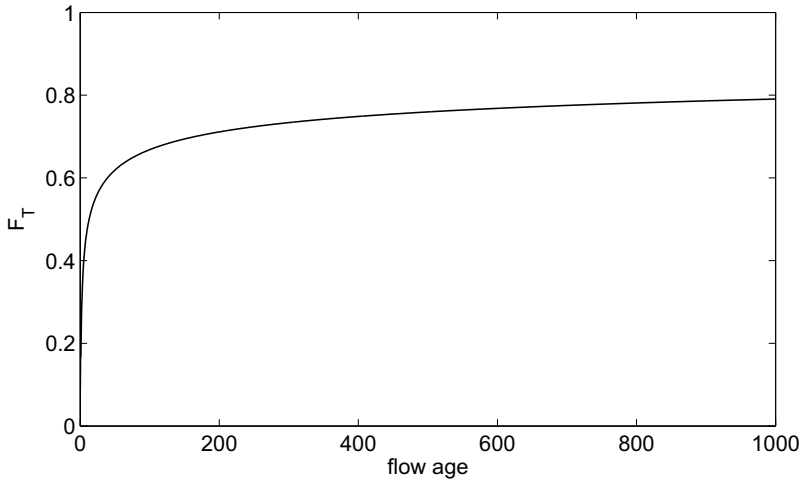

Figure 7.3: Cumulative distribution function of the flow age T, when flow lifetimes are Pareto-distributed.

density when $x < b$ comes from the first part of the piecewise defined function. In the cumulative distribution function the same appears as a constant slope, which however is hardly visible on the left edge of Figure 7.3.

7.2.2 Predictability of the renewal process

Recall that P_{11}, the probability that a flow remains open, is a function of T (Equation (7.6)). Now that the distribution of T is known, the cdf of P_{11} becomes

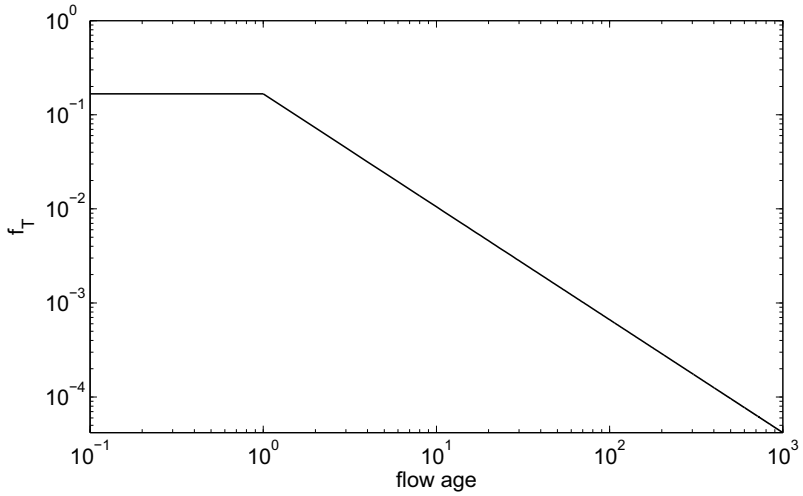

Figure 7.4: Probability density function of the flow age T on a double logarithmic scale, when flow lifetimes are Pareto-distributed.

$$
\begin{aligned}
F_P(x) &= \Pr\{P_{11} \leqslant x\} = \Pr\left\{\left(\frac{T}{T+h}\right)^a \leqslant x\right\} \\
&= \Pr\left\{T \leqslant \frac{hx^{1/a}}{1-x^{1/a}}\right\} = F_T\left(\frac{hx^{1/a}}{1-x^{1/a}}\right) \\
&= \begin{cases} \frac{(a-1)hx^{1/a}}{ab(1-x^{1/a})}, & \text{when } 0 \leqslant x < \left(\frac{b}{b+h}\right)^a \\ 1 - \frac{b^{a-1}}{a}\left(\frac{hx^{1/a}}{1-x^{1/a}}\right)^{-a+1}, & \text{when } \left(\frac{b}{b+h}\right)^a \leqslant x < 1. \end{cases}
\end{aligned} \tag{7.10}
$$

$F_P(x)$ is undefined in the unrealistic case $x = 1$, that is, when the flow would continue with probability 1. Yet, it does have the limit $F_P = 1$ when $x \to 1$.

Figure 7.5 shows the cdf of P_{11} for parameter values $a = 1.2$ and $b = 1$ and four different values of h. The curves show that with a short prediction horizon h, P_{11}

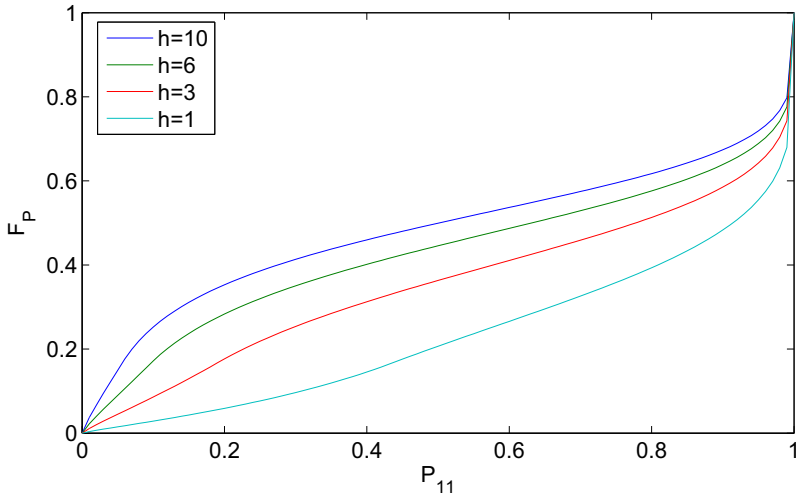

Figure 7.5: Cumulative distribution functions of the probability P_{11} with different values of the prediction horizon h.

gets relatively large values. Roughly speaking, predicting only a short time ahead implies more flows that remain open. Increasing h means trying to predict further to the future. Thus, larger values of h imply smaller probabilities because flows have more time to terminate.

The concern with the probability P_{11} has its reasons. Let us examine a network node with $N_1(t)$ active flows at time t. The expected value of the number of flows persisting until time $t+h$ is $P_{11}N_1(t)$. Furthermore, if the number of idle sources $N_0(t)$ is presumed to be known, then the expected value of $N_1(t+h)$ will be

$$\mathrm{E}\{N_1(t+h)\} = P_{11}(T,h)N_1(t) + P_{01}(T,h)N_0(t), \tag{7.11}$$

where $P_{01}(T,h)$ denotes the probability of an idle source becoming active. It is not a very realistic scenario that $N_0(t)$ be known (see Section 7.3.2), so concentrate on the term $P_{11}(T,h)N_1(t)$ of (7.11). The target is thus to predict how many of the flows currently active will be active after h time units. This will give some insight into the prediction using this approach.

If the flow durations follow the exponential distribution, (7.5) can simply be substituted into (7.11). The first term becomes $P_{11}(T,h)N_1(t) = \mathrm{e}^{-\frac{h}{\mu}}N_1(t)$, that is, the expected value is independent of the history of the flows.

But for the Pareto distribution, P_{11} is a function of T, the age of an individual connection. Because T is a random variable, also $\mathrm{E}\{N_1(t+h)\}$ becomes one. This is emphasized by denoting

$$X(T,h) = P_{11}(T,h)N_1(t) = \left(\frac{T}{T+h}\right)^a N_1(t), \tag{7.12}$$

the expected value of the number of connections remaining open after time h. Even though $X(T,h)$ is a function of t as well, argument t is omitted for simplicity.

To benefit from the heavy-tailedness of the Pareto-distributed flow lifetimes, one would like to deduce something from (7.12). The expected value of X can be calculated by [83]

$$\mathrm{E}\{X(T,h)\} = \int_{-\infty}^{\infty} X(x,h) f_T(x)\,\mathrm{d}x, \tag{7.13}$$

where $f_T(T)$ is the density function of T defined by (7.9). Note that in one sense (7.13) is an "expectation of expectation", because $X(T,h)$ is an expected value itself. The term *predictability* is introduced for $\mathrm{E}\{X(T,h)\}$ since it tells how well the number of persisting flows can be predicted in this way. The higher the predictability, the greater is the number of open flows after h time units.

On the other hand, the word predictability is somewhat questionable to describe the expected number of open flows. The process is just as predictable even if the probability is small. Rather, a high predictability should be understood as an increased confidence that the current flows will stay open. If predictability is low, the current flows are likely to terminate and new flows will possibly emerge.

Merging (7.9) and (7.12) to (7.13) yields

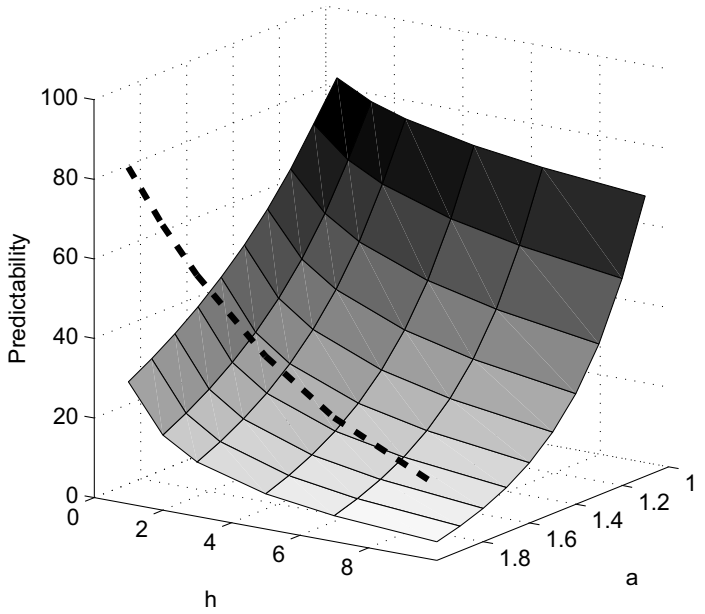

Figure 7.6: Results of numerical integration of predictability (the expected number of flows remaining open after h time units), when flow lifetimes are Pareto-distributed. The dashed line stands for the case of exponential distribution.

$$\mathrm{E}\{X(T,h)\} = N_1(t)\,\frac{a-1}{ab}\int_0^b \left(\frac{x}{x+h}\right)^a \mathrm{d}x$$
$$-N_1(t)\,\frac{(1-a)b^{a-1}}{a}\int_b^\infty \left(\frac{1}{x+h}\right)^a \mathrm{d}x. \qquad (7.14)$$

The 3D-surface in Figure 7.6 presents some results of numerical integration of Equation (7.14) within certain ranges of a and h. The Pareto location parameter had a constant value $b = 1$ in the integration, and the number of originally active flows was $N_1(t) = 100$.

As one can see from Figure 7.6, predictability of the flows decreases when h in-

creases. This is logical since predicting gets harder when the prediction horizon lengthens. The tail index a of the Pareto distribution has an even stronger influence on predictability: when a is close to 1, the flow lifetime distribution is closely related to a self-similar process (see 2.2.4) and predictability is high. But when a approaches 2, predictability descends close to zero. Generally, aggregating sources with $a \rightarrow 2$ leads to a short-range dependent process that is not predictable.

Comparing Pareto with an exponential distribution may not be fair to either one, but it is tried next. The Pareto parameter values $a = 1.2$ and $b = 1$ used above give a mean value of $ab/(a-1) = 6$. An exponential distribution with parameter $\mu = 6$ has the same mean. The dashed line in Figure 7.6 represents the predictability when flow lifetimes come from the exponential distribution with $\mu = 6$. The line does not depend on a, it is just plotted in the plane $a = 2$ for graphical clearness. Comparison shows that the Pareto distribution is clearly more predictable than the exponential when both a and h are fairly large.

7.3 Applicability of the prediction

The previous section studied the predictability of open flows when flow lifetimes follow a heavy-tailed distribution. Results showed that the expected number of flows remaining open after a certain time period is remarkably higher thanks to heavy-tailedness of the distribution, compared to an exponential distribution.

However, probabilities alone do not facilitate traffic prediction. This section discusses practical issues of applying the mathematical results to real-world traffic analysis.

The process of arriving flows can have several mathematical interpretations. Different interpretations emerge from different applications and also lead to various problems in predicting the traffic. In the following, three approaches as well as their applicability to predicting are investigated. All of the approaches are possible in a real telecommunication network.

The strength of the above predictability is that it describes the properties of the whole distribution. An individual flow need not be measured. The predictability depends only on the distribution parameters a and b, and the prediction horizon h. b, the minimum flow duration, is often obvious from the application, while a must be identified. If the age of an individual flow can be measured, then Equation (7.6) can be used to predict the future of the flow.

7.3.1 Renewal process

In Section 7.2.1, the arriving flows were interpreted as a renewal process by assuming that after a flow terminates, a new one begins immediately (Figure 7.2 a). Renewal processes of this kind are quite common in telecommunication networks, for example in a network node allocating one buffer per user. When the user finishes, the corresponding flow ends and the buffer can accommodate a new user. Examples of parallel renewal processes include the mobile station buffers in a GPRS network [14] or a peer-to-peer node with a restricted number of connections.

The predictability derived in Section 7.2.2 and plotted in Figure 7.6 tells how many of the currently active flows are expected to be active in the future. A prediction model was designed where the number of active flows at time $t+h$ could be predicted. However, the amount of active flows stays constant anyway because the model relies on renewal process theory. Thus, all that can be predicted is the proportion of current flows that remain after h time units.

Yet in some cases, finding out that a certain percentage of the current flows will persist may be useful. In a GPRS network, an arriving voice phone call may cause data connections to drop. Modelling the probability of continuing flows could help in deciding which connections to drop. This problem also relates to Quality of Service, as interrupting flows has an effect on the availability of the service. Also Call Admission Control (CAC) schemes can exploit estimates on probabilities of persisting flows. In mobile network CAC, new calls and handover calls have different probabilities [39].

Example. Consider a renewal process where a new flow begins immediately after

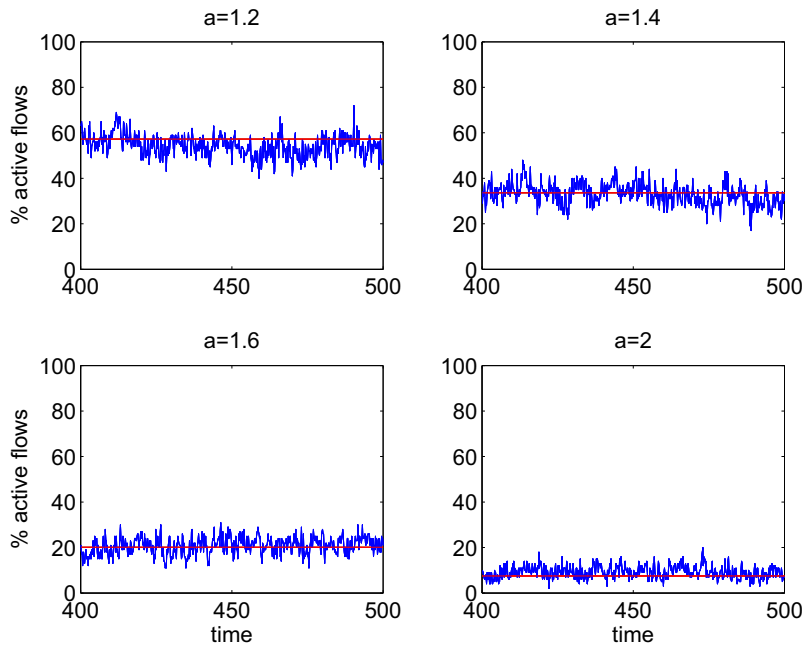

Figure 7.7: Simulated (blue) and predicted (red) flow counts of a renewal process with various Pareto shape parameter values.

the previous one terminates. The flow lifetimes are Pareto-distributed with shape parameter a. Figure 7.7 presents results from some simulations of such a process. The blue line in the graph presents the percentage of flows that remain open after $h = 6$ time units, in other words, each point on the blue line tells how many per cent of the flows are older than h. The red horizontal line is the the predicted proportion of open flows, the expected value calculated by Equation (7.14). The straightness of the red line justifies that the prediction only depends on the distribution and prediction horizon h. No individual flow history need to be measured.

Earlier in this chapter, it was assumed that the start time of the process is not known. Therefore, Figure 7.7 shows time periods after a fairly long simulation. For the same reason the four parts of the figure cannot be presented successively in the same graph. Changes in a would cause transients in the simulation — in a real process, the changes and transients would be smoother.

Comparing the four graphs in Figure 7.7 shows that the predictability of open

flows decreases along with increasing a, which is also clear from Figure 7.6. Increasing a also reduces the burstiness of the traffic, because the long-range dependence of the process decreases.

The red line predicts fairly well the average level of the blue line in each figure. This means that Equation (7.14) could be used for prediction in an application where individual flows cannot be traced but the tail index of the flow lifetime distribution can be identified. A certain amount of the currently active flows will remain active after h time units. ■

7.3.2 Alternating renewal process

If there is an idle time between two consecutive flows and also the idle times follow a probability distribution, the system is an alternating renewal process [43] (Figure 7.2 b). The distributions of the active and idle times need not be similar. A familiar example of parallel alternating renewal processes is a GSM network cell with free capacity. Each time slot can hold a call, and after the call there is a pause before the next one. An interesting application of heavy-tailed distributions would be to predict the number of active flows at a random time instant.

Equation 7.11 describes $N_0 + N_1$ parallel alternating renewal processes. At each time instant, the N_1 active flows form a sample of renewal processes whose residual lifetimes can be modelled. Likewise, the model tells how long the N_0 idle sources will probably be idle and thus gives an estimate of the number of active flows in the future.

In network traffic modelling, it is common to model the active (on) times with Pareto and idle (off) times with exponential distribution. Unfortunately, Equation (7.7) does not hold when the on and off times are not equally distributed. The equation can be used only if F_D is replaced by a cdf that treats the cycle of one flow and the subsequent off time as one renewal. In this case, the pdf (and further cdf) of the on-off-cycle could be found through convolution of the on and off times:

$$f_{\text{on}}(t) * f_{\text{off}}(t) = \int_0^t f_{\text{on}}(t-s) f_{\text{off}}(s) \mathrm{d}s \tag{7.15}$$

However, this method treats the combined on-off-cycle as one renewal. [96] The convoluted pdf no longer gives information of a single flow from its beginning to end; it concerns only the time from the beginning of a flow to the beginning of the next flow. Thus, the methodology presented in this chapter is not usable in alternating renewal processes with differently distributed on and off times.

Assuming on and off times equally distributed is probably not far from the truth in most cases. This makes it possible to predict the N_1 active sources and N_0 idle sources separately, similarly to the above example. Both active sources remaining active and idle sources remaining idle can be predicted. Yet, another drawback is that N_0 is not necessarily known.

7.3.3 No renewals

Even a single user or application acts as an alternating renewal process in the network. Flows caused by the user follow a probability distribution, while the idle times between flows may follow another distribution. These flows accumulate at a router, but the router does not see a number of renewal processes. Instead, the data arrive at the router in packets, and the router serves its clients in packets, not in flows. Thus, a router working at its full capacity might be serving only one single flow with lots of data. If, on the other hand, several clients send packets simultaneously, the packets and thus flows are multiplexed and several flows are active.

For the above reasons, modelling traffic at router level is difficult with the presented approach. Other possibilities may be found by examining the packet size distributions, which however can never be quite pure since there are limits for the packet size set by different components of the network.

7.4 Conclusion

If flow durations follow a heavy-tailed distribution, a flow is the more likely to remain open the longer it has been open. This fact was expanded and developed in this chapter towards predictability of traffic volume. However, the suitability of the prediction model to practical applications remained limited.

Under certain conditions, the flows constitute a renewal process, and the requirements for prediction are fulfilled. Examples of renewal processes in telecommunications include mobile station buffers, voice calls and peer-to-peer nodes. But in many widely studied applications, such as routers, the presented prediction model is hardly applicable.

Chapter 8

Conclusion

The starting point of this thesis was the heavy-tailedness of network traffic. Several studies have shown that the traffic in telecommunication networks contains heavy-tailed distributions. Some of the reported findings concentrate on the heavy tail and its tail index [28], while some speak of self-similarity [87, 38]. The two phenomena have a close relationship: heavy tails cause self-similarity [117].

Traditionally, heavy-tailed distributions have been regarded as a handicap to both network dimensioning and data analysis. The target of this work was to provide a set of tools for facilitating some typical tasks of network traffic data analysis. Although the tools, the chapters of the thesis, were fairly separate entities, their common features included statistical testing and change detection in addition to heavy-tailed distributions. This final chapter summarizes the results achieved and contemplates the meaning of them as well as the entirety of the work.

8.1 Results

The thesis first discussed the use of conventional goodness-of-fit tests with real network data. Two data sets, one of which originates from a famous study in the 1990's [26], were tested for known distributions. One data set showed some visual

resemblance to lognormal distribution, the other had a heavy tail according to the original publication. Nevertheless, neither of the sets passed a goodness-of-fit test.

The impure distributions of real-world data are perhaps a reason for the popularity of various visualization methods. Stringent goodness-of-fit tests do not necessarily reveal anything, but human eye can recognize even a weak pattern. On the other hand, visualization has its limits, which will be discussed later in this chapter.

Chapter 4 started from the fact that the data are often compressed into a histogram. The traditional χ^2 test was first found eligible for detecting changes in, for example, port number histograms. Then a test scenario was created where continuous data were converted into a histogram and analyzed as if they had come from a network management system as a histogram. Even though the data did not contain actual heavy-tailed distributions, the histogram was quite sparse, which is typical to the peculiarities of network traffic data. Gaps in a histogram make many popular tests useless, thus Monte Carlo simulation was adopted for change detection. The method proved to find well some invisible changes, while some other changes could not be detected.

Standard tests do not work well with traffic data, as was stated in Chapter 3. Consequently, Chapter 4 introduced Monte Carlo simulation for designing more advanced tests. Kolmogorov-Smirnov statistic was used with little consideration even though it is known to be an inferior choice for weird distributions. Hence, Chapter 5 investigated which test statistic would work best with network traffic data. The results were quite surprising: Many statistics, though introduced in the literature together with heavy-tailed distributions, did not survive change detection between two different traffic types. Most of the statistics measured dispersion, and thus one single outlier was enough to confuse them. For the same reason, the plain median beats the more complicated statistics.

Chapter 6 went into the subject that acted in the background also in earlier chapters: sample size. Little attention has been paid to excessive sampling in statistical testing, as opposed to sufficient sampling. The sample size study in Chapter 6 showed that despite the huge data reservoir, a relatively small sample of $n = 100$ seemed suitable for change detection. When the null hypothesis is exactly correct, the sample size has no effect. In practical data analysis however, the data

never perfectly match the hypothesis; the question is about significance. This is particularly true for change detection.

Heavy-tailed distributions were the core of the thesis. Chapter 7 explored an interesting property of heavy-tailedness, that is, predictability. It was fairly easy to show that if a flow length follows a heavy-tailed distribution, the flow is the more likely to remain open the longer it has been open. This property was further developed into a prediction model. The results showed that under certain circumstances, heavy-tailed distributions might really be useful in predicting. Yet, the assumptions made in the model limit its applicability to some relatively rare cases.

8.2 Discussion

The title of this thesis contains the term *data analysis*. Berthold and Hand define data analysis briefly as "processing of those data" [17], which certainly covers all methods presented in the thesis.

Data mining in turn is a newish term that emphasizes the large amount of data and covers more tasks than just data analysis [46]. This thesis contained many features typical to data mining according to [46]. The data sets were analyzed off-line, and one of the assumptions was that the analyst cannot influence the data collection infrastrucure but uses data provided as is.

On the other hand, the thesis developed most of the methods with online monitoring in mind. Even though the test cases used mostly off-line data, the change detection methods suit well for online monitoring. This takes the thesis closer to *process management*, which Alhoniemi divides in monitoring and analysis of process data [5]. Indeed, a telecommunication network should be viewed as an industrial process even though it lacks natural phenomena.

The above discussion shows that defining or classifying any methods into a strict category is hardly possible. Therefore, let us review the contents of this thesis from a functional point of view.

Visualization is an essential part of exploratory data analysis. The information content of the data should be compressed into a form detectable by human eye. While extremely useful for many purposes, visualization gets inconvenient when data volumes explode. Even human eye cannot detect thousands of figures.

This thesis seeks for an utmost form of data visualization. Think of the network as a process. The process has a control room, where an operator supervises the process. An experienced operator can perceive the state of the process from several pieces of information, yet a single indicator telling when something goes wrong — a red light — would be desirable. A concept where a process monitoring method has a central role is presented in [91].

The change detection methods presented in this thesis aim at implementation of the red light. Chapters 3–6 all dealt with various aspects of change detection that are useful in process monitoring. Chapter 7 went one step further; it tried to predict traffic volume, which, too, is useful in monitoring. It would undoubtedly be useful to switch the red light on before something goes wrong.

This thesis, as many other studies, uncovered the everlasting truth of data analysis: No single method can be blindly applied to all data. Even though the methods were developed with online monitoring in mind, various choices — such as significance limits, bin widths or censoring thresholds — had to be made before the setout. Additionally; outliers, breaks and natural variation always defeat parts of the data, and preprocessing phases as well as human expertise are needed before the actual analysis. In other words, statistical methods do not reduce the importance of preprocessing.

In textbook examples, stationarity of the process is usually presumed without further consideration. Yet this thesis paid little attention to the stationarity of the data sets. This was because change detection was one of the main goals of the thesis, and stationary processes have no changes. Hence, change detection with goodness-of-fit tests is always a compromise. If there is a slow ongoing change in the process, the sample collected will necessarily contain part of the change. Whether or not the detection method can distinguish this sample from the next one, depends on many details. Detection of abrupt changes is easier; now the process can be considered stationary until a change occurs.

The cases presented in this thesis represented relatively specific applications of network traffic data analysis. For example, the power studies in Chapters 5 and 6 embodied change detection between HTTP and Gnutella traffic only. As explained in Chapter 5, these two protocols were chosen to represent a typical case in network traffic, but certainly other cases exist. Furthermore, Chapter 7 concentrated on a certain scenario of traffic prediction: It resorted to renewal processes and number of flows. Other prediction mechanisms based on flow duration heavy-tailedness may exist.

Bibliography

[1] Inmaculada B. Aban and Mark M. Meerschaert. Generalized least-squares estimators for the thickness of heavy tails. *Journal of Statistical Planning and Inference*, 119(2):341–352, February 2004.

[2] Inmaculada B. Aban, Mark M. Meerschaert, and Anna K. Panorska. Parameter estimation for the truncated Pareto distribution. *Journal of the American Statistical Association*, 101(473), March 2006.

[3] Patrice Abry and Darryl Veitch. Wavelet analysis of long-range-dependent traffic. *IEEE Transactions on Information Theory*, 44(1), January 1998.

[4] Abdelnaser Adas. Traffic models in broadband networks. *IEEE Communications Magazine*, 3(7):82–89, 1997.

[5] Esa Alhoniemi. *Unsupervised Pattern Recognition Methods for Exploratory Analysis of Industrial Process Data*. PhD thesis, Helsinki University of Technology, Finland, 2002.

[6] The American Heritage® Dictionary of the English language. `http://www.thefreedictionary.com/`. Retrieved June 10, 2009.

[7] Nikkie C. Anand, Caterina Scoglio, and Balasubramaniam Natarajan. GARCH - non-linear time series model for traffic modeling and prediction. In *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, 2008.

[8] T. W. Anderson and D. A. Darling. A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–679, December 1954.

[9] Mauro Andreolini, Sara Casolari, and Michele Colajanni. Models and framework for supporting runtime decisions in web-based systems. *ACM Transactions on the Web*, 2(3):1–43, 2008.

[10] Peter Armitage. The design and analysis of clinical trials. In S. Ghosh and C. R. Rao, editors, *Handbook of Statistics*, volume 13, chapter 1, pages 1–29. Elsevier Science B. V., 1996.

[11] Lee J. Bain and Max Engelhardt. *Statistical analysis of reliability and life-testing models: theory and methods*. Marcel Dekker, Inc., second edition, 1991.

[12] J. Wesley Barnes. *Statistical Analysis for Engineers and Scientists*. McGraw-Hill, 1994.

[13] Yuliy Baryshnikov, Ed Coffman, Guillaume Pierre, Dan Rubenstein, Mark Squillante, and Teddy Yimwadsana. Predictability of web-server traffic congestion. In *Proceedings of the 10th International Workshop on Web Content Caching and Distribution (WCW'05)*, 2005.

[14] Regis J. (Bud) Bates. *GPRS*. McGraw-Hill, 2002.

[15] Jan Beran. *Statistics for Long-Memory Processes*. Chapman & Hall/CRC, 1994.

[16] Jan Beran, Robert Sherman, Murad S. Taqqu, and Walter Willinger. Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, 43(2):1566–1579, 1995.

[17] Michael Berthold and David J. Hand, editors. *Intelligent Data Analysis: an introduction*. Springer, 1999.

[18] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[19] M. T. Boswell, S. D. Gore, G. P. Patil, and C. Taillie. The art of computer generation of random variables. In C. R. Rao, editor, *Handbook of Statistics*, volume 9, chapter 20, pages 661–721. Elsevier Science Publishers, 1993.

[20] M. F. Brilhante. Exponentiality versus generalized Pareto — a resistant and robust test. *REVSTAT — Statistical Journal*, 2(1), 2004.

[21] Marek Bromirski and Wieslaw Lobejko. Fractals and chaos for modelling multimedia ATM traffic. In Demetres Kouvatsos, editor, *Performance Evaluation and Applications of ATM Networks*, pages 31–50. Kluwer Academic Publishers, 2002.

[22] Nevil Brownlee and KC Claffy. Understanding Internet traffic streams: dragonflies and tortoises. *IEEE Communications Magazine*, 40(10):110–117, October 2002.

[23] V. Choulakian and M. A. Stephens. Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4), 2001.

[24] Cisco IOS NetFlow. `http://www.cisco.com/go/netflow/`. Retrieved June 10, 2009.

[25] W. J. Conover. A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association*, 67(339):591–596, 1972.

[26] Mark Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. In *Proceedings of SIGMETRICS'96: The ACM International Conference on Measurement and Modeling of Computer Systems*, 1996.

[27] Mark E. Crovella and Lester Lipsky. Simulations with heavy-tailed workloads. In Kihong Park and Walter Willinger, editors, *Self-similar network traffic and performance evaluation*, chapter 3. John Wiley & Son's, Inc., 2000.

[28] Mark E. Crovella and Murad S. Taqqu. Estimating the heavy tail index from scaling properties. *Methodology and Computing in Applied Probability*, 1(1):1–22, 1999.

[29] Mark E. Crovella, Murad S. Taqqu, and Azer Bestavros. Heavy-tailed probability distributions in the World Wide Web. In Robert J. Adler, Raisa E. Feldman, and Murad S. Taqqu, editors, *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Birkhäuser, 1998.

[30] Carlos R. Cunha, Azer Bestavros, and Mark E. Crovella. Characteristics of WWW client-based traces. Technical report, Computer Science Department, Boston University, 1995.

[31] Ralph B. D'Agostino and Michael A. Stephens. Overview. In Ralph B. D'Agostino and Michael A. Stephens, editors, *Goodness-of-fit techniques*, chapter 1. Marcel Dekker, Inc., 1986.

[32] Luc Devroye. *Non-uniform random variate generation*. Springer-Verlag, 1986.

[33] Luc Devroye. *A Course in Density Estimation*. Birkhäuser, 1987.

[34] Stephen Dill, Ravi Kumar, Kevin S. McCurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology*, 2(3):205–223, August 2002.

[35] Edward R. Dougherty. *Probability and statistics for the engineering, computing, and physical sciences*. Prentice-Hall, 1990.

[36] N. G. Duffield, J. T. Lewis, Neil O'Connell, Raymond Russell, and Fergal Toomey. Statistical issues raised by the bellcore data. In *Proceedings of the 11th Teletraffic Symposium*, 1994.

[37] Lasse M. Eriksson and Mikael Johansson. PID controller tuning rules for varying time-delay systems. In *Proceedings of the 2007 American Control Conference*, pages 619–625, 2007.

[38] Ashok Erramilli, Matthew Roughan, Darryl Veitch, and Walter Willinger. Self-similar traffic and network dynamics. *Proceedings of the IEEE*, 90(5):800–819, May 2002.

[39] Yuguang Fang and Yi Zhang. Call admission control schemes and performance analysis in wireless mobile networks. *IEEE Transactions on Vehicular Technology*, 51(2):371–382, 2002.

[40] Roger L. Freeman. *Telecommunication system engineering*. John Wiley & Sons, Inc., 2004.

[41] `ftp://cs-ftp.bu.edu/techreports/95-010-web-client-traces.tar.gz`. Retrieved September 27, 2006.

[42] Mark W. Garrett and Walter Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *Proceedings of ACM SIGCOMM'94*, pages 269–280, 1994.

[43] Ilya B. Gertsbakh. *Statistical reliability theory*. Marcel Dekker, Inc., 1989.

[44] Per-Erik Hagmark. Tilastomatemaattinen datan käsittely ja näytteen otto. Lecture notes, Tampere University of Technology, 2006.

[45] Hassan Hajji. Statistical analysis of network traffic for adaptive faults detection. *IEEE Transactions on Neural Networks*, 16(5), September 2005.

[46] David Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

[47] Guanghui He, Yuan Gao, Jennifer C. Hou, and Kihong Park. A case for exploiting self-similarity of network traffic in TCP congestion control. *Computer Networks*, 45(6):743–766, 2004.

[48] Félix Hernández-Campos, J. S. Marron, Gennady Samorodnitsky, and F. D. Smith. Variable heavy tails in Internet traffic. *Performance Evaluation*, 58:261–284, 2004.

[49] Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, September 1975.

[50] Melvin J. Hinich and Robert E. Molyneux. Predicting information flows in network traffic. *Journal of the American society for information science and technology*, 54(2):161–168, 2003.

[51] Susan Dadakis Horn. Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, 33(1):237–247, March 1977.

[52] Polly Huang, Anja Feldmann, and Walter Willinger. A non-intrusive, wavelet-based approach to detecting network performance problems. In *Proceedings of ACM/SIGCOMM Internet Measurement Workshop 2001*, 2001.

[53] Dragos Ilie, David Erman, Adrian Popescu, and Arne A. Nilsson. Measurement and analysis of Gnutella signaling traffic. In *Proceedings of the International IPSI-2004 Conference*, 2004.

[54] Internet Assigned Numbers Authority (IANA). Port numbers. `http://www.iana.org/assignments/port-numbers`. Retrieved June 10, 2009.

[55] L. Janowski, T. Ziegler, and E. Hasenleithner. A scaling analysis of UMTS traffic. In *Proceedings of the Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN'06)*, 2006.

[56] Michel C. Jeruchim, Philip Balaban, and K. Sam Shanmugan. *Simulation of Communication Systems*. Kluwer Academic Publishers, 2000.

[57] Sung-Don Joo, Chae-Woo Lee, and Yeon Hwa Chung. Analysis and modeling of traffic from residential high speed Internet subscribers. In *Lecture Notes in Computer Science*, pages 410–419. Springer, 2004.

[58] Thomas Karagiannis, Mart Molle, and Michalis Faloutsos. Long-range dependence. Ten years of Internet traffic modeling. *IEEE Internet Computing*, 8(5):57–64, 2004.

[59] Bernhard Klar. Goodness-of-fit tests for discrete models based on the integrated distribution function. *Metrika*, 49(1):53–69, 1999.

[60] Glen Kramer, Biswanath Mukherjee, and Gerry Pesavento. Interleaved polling with adaptive cycle time (IPACT): A dynamic bandwidth distribution scheme in an optical access network. *Photonic Network Communications*, 4(1):89–107, 2001.

[61] Balachander Krishnamurthy, Harsha V. Madhyastha, and Suresh Venkatasubramanian. On stationarity in Internet measurements through an information-theoretic lens. In *Proceedings of the 21st International Conference on Data Engineering (ICDE '05)*, 2005.

[62] James F. Kurose and Keith W. Ross. *Computer networking: a top-down approach featuring the Internet*. Pearson Education, Inc., third edition, 2005.

[63] Mikko Laurikkala, Tapani Honkanen, and Hannu Koivisto. Elephant flows snatch a lion's share of network capacity. In *Proceedings of the Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN'04)*, 2004.

[64] Mikko Laurikkala and Hannu Koivisto. Power study of statistical tests for network traffic change detection. In *Proceedings of the 30th International Conference on Information Technology Interfaces (ITI2008)*, 2008.

[65] Mikko Laurikkala and Hannu Koivisto. Test statistics and sample sizes in network traffic change detection. *Computational Statistics and Data Analysis*, 2009. Submitted.

[66] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic. In *Proceedings of ACM SIGCOMM'93*, pages 183–193, 1993.

[67] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1), February 1994.

[68] Will E. Leland and Daniel V. Wilson. High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection. In *INFOCOM '91. Proceedings of the Tenth Annual Joint Conference of the IEEE Computer and Communications Societies*, 1991.

[69] Russell V. Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193, 2001.

[70] Lundy Lewis. *Service level management for enterprise networks*. Artech House, 1999.

[71] Irene C. Y. Ma and James Irvine. Characteristics of WAP traffic. *Wireless Networks*, 10(1):71–81, 2004.

[72] Benoit B. Mandelbrot. *Fractals: Form, Chance and Dimension*. W. H. Freeman and Company, San Fransisco, 1977.

[73] Constantine Manikopoulos and Symeon Papavassiliou. Network intrusion and fault detection: A statistical anomaly approach. *IEEE Communications Magazine*, 40(10):76–82, October 2002.

[74] Dimitris G. Manolakis, Vinay K. Ingle, and Stephen M. Kogon. *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering, and Array Processing*. Artech House, Inc., 2005.

[75] J. S. Marron, F. Hernandez-Campos, and F. D. Smith. Mice and elephants visualization of network traffic. In *Proceedings of Compstat 2002*, 2002.

[76] Don L. McLeish. *Monte Carlo simulation and finance*. John Wiley & Sons, Inc., 2005.

[77] John R. Michael and William R. Schucany. Analysis of data from censored samples. In Ralph B. D'Agostino and Michael A. Stephens, editors, *Goodness-of-fit techniques*, chapter 11. Marcel Dekker, Inc., 1986.

[78] J. S. Milton and Jesse C. Arnold. *Introduction to Probability and Statistics*. McGraw-Hill, second edition, 1990.

[79] David S. Moore. Tests of chi-squared type. In Ralph B. D'Agostino and Michael A. Stephens, editors, *Goodness-of-fit techniques*, chapter 3. Marcel Dekker, Inc., 1986.

[80] Ilkka Norros. On the use of fractional brownian motion in the theory of connectionless networks. *IEEE Journal on Selected Areas in Communications*, 13(6):953–962, August 1995.

[81] Sven A. M. Östring and Harsha Sirisena. The influence of long-range dependence on traffic prediction. In *Proceedings of IEEE International Conference on Communications*, 2001.

[82] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, second edition, 1991.

[83] Athanasios Papoulis and S. Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. McGraw-Hill, fourth edition, 2002.

[84] Kihong Park and Walter Willinger. Self-similar network traffic: an overview. In *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons, Inc., 2000.

[85] Peter F. Pawlita. Traffic measurements in data networks, recent measurement results, and some implications. *IEEE Transactions on Communications*, 29(4):525–535, April 1981.

[86] Vern Paxson. Empirically-derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking*, 2(4):316–336, August 1994.

[87] Vern Paxson and Sally Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[88] Vitaly Petroff. Self-similar network traffic: From chaos and fractals to forecasting and QoS. In *Proceedings of the Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN'04)*, 2004.

[89] Leo Pipino, Richard Wang, David Kopcso, and William Rybolt. Developing measurement scales for data-quality dimensions. In Richard Y. Wang, Elizabeth M. Pierce, Stuart E. Madnick, and Craig W. Fisher, editors, *Information quality*, chapter 3. M.E. Sharpe, Inc., 2005.

[90] Principles for a telecommunications management network. ITU-T Recommendation M.3010, 2000.

[91] Pietari Pulkkinen, Mikko Laurikkala, Aino Ropponen, and Hannu Koivisto. Quality management in GPRS networks with fuzzy case-based reasoning. *Knowledge-Based Systems*, 21(5):421–428, 2008.

[92] C. P. Quesenberry. Some transformation methods in goodness-of-fit. In Ralph B. D'Agostino and Michael A. Stephens, editors, *Goodness-of-fit techniques*, chapter 6. Marcel Dekker, Inc., 1986.

[93] P. Révész. Density estimation. In P. R. Krishnaiah and P. K. Sen, editors, *Handbook of Statistics*, volume 4, chapter 24, pages 531–549. Elsevier Science Publishers, 1984.

[94] Sheldon M. Ross. *Stochastic Processes*. John Wiley & Sons, Inc., second edition, 1996.

[95] Matthew Roughan, Darryl Veitch, and Patrice Abry. Real-time estimation of the parameters of long-range dependence. *IEEE/ACM Transactions on Networking*, 8(4):467–478, August 2000.

[96] Keijo Ruohonen. Luotettavuus, käytettävyys, huollettavuus. Lecture notes, Tampere University of Technology, 2002.

[97] Mikko Salmenperä and Jari Seppälä. Data security considerations in modern automation networks. In *1st international conference on informatics in control, automation and robotics (ICINCO 2004)*, 2004.

[98] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[99] Khushboo Shah, Stephan Bohacek, and Edmond Jonckheere. On the predictability of data network traffic. In *Proceedings of the 2003 American Control Conference*, 2003.

[100] Jun Shao. *Mathematical Statistics*. Springer-Verlag, 1999.

[101] Oleg I. Sheluhin, Sergey M. Smolskiy, and Andrey V. Osin. *Self-Similar Processes in Telecommunications*. John Wiley & Sons, Ltd., 2007.

[102] Jonathan J. Shuster. *CRC handbook of sample size guidelines for clinical trials*. CRC Press, 1990.

[103] V. Kerry Smith. A simulation analysis of the power of several tests for detecting heavy-tailed distributions. *Journal of the American Statistical Association*, 70(351):662–665, September 1975.

[104] Murray R. Spiegel, John J. Schiller, and R. Alu Srinivasan. *Probability and Statistics*. McGraw-Hill, second edition, 2000.

[105] Michael C. Steele. *The power of categorical goodness-of-fit test statistics*. PhD thesis, Griffith University, Australia, 2002.

[106] Michael A. Stephens. Tests based on EDF statistics. In Ralph B. D'Agostino and Michael A. Stephens, editors, *Goodness-of-fit techniques*, chapter 4. Marcel Dekker, Inc., 1986.

[107] W. Richard Stevens. *TCP/IP Illustrated, Volume 1: The Protocols*. Addison-Wesley, 1994.

[108] Mani Subramanian. *Network management: An introduction to principals and practice*. Addison-Wesley, 2000.

[109] Andrew S. Tanenbaum. *Computer Networks*. Pearson Education, Inc., fourth edition, 2003.

[110] Murad S. Taqqu, Vadim Teverovsky, and Walter Willinger. Estimators for long-range dependence: an empirical study. *Fractals*, 3(4):785–788, 1995.

[111] Murad S. Taqqu, Walter Willinger, and Robert Sherman. Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review*, 27(2), April 1997.

[112] Henry C. Thode, Jr. *Testing for Normality*. Marcel Dekker, Inc., 2002.

[113] Marina Thottan and Chuanyi Ji. Statistical detection of enterprise network problems. *Journal of Network and Systems Management*, 7(1):27–45, 1999.

[114] Timothy C. Urdan. *Statistics in Plain English*. Lawrence Erlbaum Associates, Inc., 2005.

[115] M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51(1), 1997.

[116] Xin Wang and Xiuming Shan. A wavelet-based method to predict Internet traffic. In *International Conference on Communications, Circuits and Systems and West Sino Expositions, IEEE*, 2002.

[117] Walter Willinger, Murad S. Taqqu, Robert Sherman, and Daniel V. Wilson. Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, 1997.

[118] Constance L. Wood and Michele M. Altavela. Large-sample results for Kolmogorov-Smirnov statistics for discrete distributions. *Biometrika*, 65(1):235–239, 1978.

[119] Ge Xiaohu, Shaokai Yu, Won-Sik Yoon, and Yong-Deak Kim. A new prediction method of alpha-stable processes for self-similar traffic. In *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, 2004.

[120] Nong Ye, Syed Masum Emran, and Sean Vilbert. Multivariate statistical analysis of audit trails for host-based intrusion detection. *IEEE Transactions on Computers*, 51(7), July 2002.

[121] Robert H. Zakon. Hobbes' Internet timeline. `http://www.zakon.org/robert/internet/timeline/`. Retrieved June 10, 2009.