UNIVERSITÄT ROSTOCK

Diplomarbeit

Vergleichende Betrachtungen statistischer Verfahren zur Beschreibung von Datenverkehr in IP Netzen

vorgelegt am : 05. April 2005

von : cand. ing. Jens Kosubek
A.-J.-Krusensternstraße 32
18106 Rostock

Matrikelnummer : 098201860

Betreuer : Dr.-Ing. Hans-Dietrich Melzer
Dipl.-Ing. Thomas Kessler

Aufgabenstellung

Vergleichende Betrachtungen statistischer Verfahren zur Beschreibung von Datenverkehr in IP Netzen

Die Analyse realer Kommunikationsaufkommen dient unter Anderem der Optimierung von Kommunikationsnetzwerken und Datenströmen.
Herr Kosubek soll im Rahmen der Diplomarbeit aufbauend auf vorangegangene Arbeiten die Aufzeichnung von Daten im Netz des Institutes für Nachrichtentechnik und Informationselektronik sowie im ComLab unter spezieller Berücksichtigung von Zeitpunkt und Dauer der Messungen optimieren und Aufzeichnungen durchführen.
Um zu aussagefähigen Ergebnissen zu gelangen sind die Wahrscheinlichkeitsdichteverteilungsfunktionen von Zwischenankunftszeiten und Paketlängen zu bestimmen und durch aus der Literatur bekannte Verteilungen (z.B. die Paretoverteilung und die Weibullverteilung) zu beschreiben. Dazu sind durch den Kandidaten Qualitätskriterien für Schätzungen und den Vergleich der Verfahren aufzustellen.
Die Parameterschätzungen sind in Hinblick auf die Nutzung zur Signalsynthese z.B. für die Generierung von Testfolgen zwecks Netzwerksimulation zu untersuchen und unter OPNET zu testen.

Zusammenfassung

Vergleichende Betrachtungen statistischer Verfahren zur Beschreibung von Datenverkehr in IP Netzen

Aussagen über das zukünftige Verhalten von Netzwerkverkehr können über entsprechende Simulationen erfolgen. Dazu werden Verkehrsmodelle benötigt, die das real auftretende Datenverkehrsaufkommen zuverlässig und genau beschreiben.
Die Auswahl von geeigneten Modellen ist eng mit den darunterliegenden Wahrscheinlichkeitsverteilungen verbunden, mit denen die Wahrscheinlichkeit des Auftretens eines Ereignisses wiedergegeben werden kann.
Im Rahmen der vorliegenden Diplomarbeit werden unterschiedliche Verteilungen genauer betrachtet in Hinblick auf deren Möglichkeiten der Beschreibung von Netzwerkverkehrsparametern, wie den Zwischenankunftszeiten und Paketlängen. Besondere Bedeutung kommt hier den Verteilungen zu, die über die Eigenschaft der „heavy-tailedness“ verfügen.
Detailliert dargestellt werden statistische Verfahren zur Bestimmung der Verteilungsparameter. Aufgezeigt werden dabei deren Vor- und Nachteile, in Bezug auf die Genauigkeit der Schätzung.
Weiterhin werden Gütekriterien untersucht, mit deren Hilfe die Anpassung zwischen empirischer und theoretischer Verteilungsfunktion quantitativ verglichen werden kann. Betrachtet werden sowohl klassische Anpassungstests als auch alternative Kriterien. Anschließend erfolgt die Anwendung der theoretischen Betrachtungen auf mehrere Meßreihen.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AKF	Autokorrelationsfunktion
ARIMA	Auto Regressive Integrated Moving Average
CDF	Cumulative Distribution Function
CPU	Central Processing Unit
DDR	Double Data Rate
FARIMA	Fractional ARIMA
FGN	Fractional Gaussian Noise
FOTS	Fractal Onset Time Scale
FPGA	Field Programmable Gate Array
GbE	Gigabit Ethernet
HTTP	Hypertext Transfer Protocol
ICMP	Internet Control Message Protocol
IP	Internet Protocol
ISO	International Organization for Standardization
K-S-Test	Kolmogorov-Smirnov Test
LAN	Local Area Network
LRD	Long-Range Dependence
LSE	Least Squares Estimator
MLE	Maximum Likelihood Estimator
MSE	Mean Square Error
NIC	Network Interface Card
OSI	Open Systems Interconnection
P-P-Plot	Probability-Probability-Plot
PCI	Peripheral Component Interface

Q-Q-Plot	Quantile-Quantile-Plot
RAM	Random Access Memory
RMS	Root Mean Square Error
RPG	Raw Packet Generator
SQL	Structured Query Language
SRD	Short-Range Dependence
TCP	Transfer Control Protocol
UNIX	Uniplexed Information and Computing System
VTP	Variance-Time-Plot
WAN	Wide Area Network

Symbolverzeichnis

lateinische Symbole

a	Parameter der Geradengleichung $y = ax + b$
b	Parameter der Geradengleichung $y = ax + b$
d	Vektor der aboluten Differenzen $d = \|F_X(x) - F_0(x)\|$
d_{max}	maximale abolute Differenz $d_{max} = max\,\|F_X(x) - F_0(x)\|$
EX	Erwartungswert (auch: $E[X]$)
$f(\cdot)$	Dichtefunktion einer Zufallsvariablen
$F(\cdot)$	Verteilungsfunktion einer Zufallsvariablen
$F_0(x)$	theoretische Verteilungsfunktion
$F_X(x)$	empirische Verteilungsfunktion
H	Hurst-Parameter
H_0	Nullhypothese wird angenommen, wenn gilt $H_0 : F_X(x) = F_0(x\|\hat{\theta})$
H_1	Nullhypothese wird abgelehnt, wenn gilt $H_1 : F_X(x) \neq F_0(x\|\hat{\theta})$
$I(\cdot)$	Periodogramm
$\frac{R(n)}{S(n)}$	Rescaled Adjusted Range Statistik
M_k	k-te Moment einer Zufallsvariablen X
r	Korrelationskoeffizient
$r(k)$	Autokorrelationsfunktion
t_0	Lageparameter der Paretoverteilung
$Var\ X$	Varianz

kalligraphische Symbole

$\mathcal{L}(x\|\theta)$	Likelihood-Funktion der Beobachtungen $[x1, \ldots, x_N]$ mit den Parametern $[\theta_1, \ldots, \theta_k]$
$\mathcal{N}(\mu, \sigma)$	Normalverteilung mit den Parametern μ und σ

griechische Symbole

α	Signifikanzniveau
α_p	Formparameter der Paretoverteilung
α_w	Formparameter der Weibullverteilung
β	Vergrößerungsparameter der Weibullverteilung
γ	Lageparameter der Weibullverteilung
$\Gamma(\cdot)$	Gammafunktion mit: $\Gamma(x) = \int_{x=0}^{\infty} t^{x-1}e^{-t}dt$ für $x > 0$.
λ	Verteilungsparameter der Exponentialverteilung
Λ	log-likelihood Funktion $\Lambda = ln(\mathcal{L})$
μ	Lageparameter der Normalverteilung und arithmetisches Mittel
μ_L	Skalierungsparameter der Lognormalverteilung
σ	Skalierungsparameter der Normalverteilung und Standardabweichung
σ_L	Formparameter der Lognormalverteilung
θ	Parameter einer Verteilung
$\hat{\theta}$	geschätzter Parameter

Weitere Symbole werden im entsprechenden Kontext erläutert.

Kapitel 1

Einleitung

1.1 Motivation und Gliederung der Arbeit

Ein Ziel bei der Planung und dem Management von paketorientierten Netzwerken liegt in der effizienten Auslastung der zugrundeliegenden Infrastruktur.
Hierzu werden geeignete Modelle benötigt, die auf Ankunftsprozessen basieren. Der Begriff Prozess ist allgemein als Vorgang definiert. Speziell mit Ankunftsprozessen wird das Verhalten von Ankünften (z.B. von Paketen) beschrieben. Der eigentliche Vorgang ist hier das Warten auf das nächste Eintreffen eines Pakets (vgl. (Nie04)).
In der Vergangenheit wurden die Ankunftsprozesse mit Poissonmodellen beschrieben, die auf Gesetzmäßigkeiten der leitungsvermittelten Telefonie beruhen. Es wurde davon ausgegangen, daß die Ankünfte strikt voneinander unabhängig und daß die Zwischenankunftszeiten immer exponentiell verteilt sind, mit nur einem Parameter λ, der mittleren Ankunftsrate. Die leitungsvermittelte Telefonie geht auch davon aus, daß die Wachstumsraten vorhersagbar sind und daß jedes Vorgehen in den Netzwerken zentral verwaltet wird (vgl. (WP98)).
Mit Beginn der Übertragung von Daten hat sich das jedoch geändert. Die Charakteristika einzelner Verbindungen zeigten sich als extrem variabel und so schwankte zum Beispiel die Verbindungsdauer zwischen extrem kurz und extrem lang. In einigen Veröffentlichungen wird das Verhalten als „chaotisch“ beschrieben. Wie aus (Pax94) zu entnehmen ist,

waren die in den 80er und frühen 90er Jahren durchgeführten Studien zum Thema Netzwerkverkehr unzureichend. Die betrachteten Verteilungen wurden nur mittels ihrer ersten beiden Momente charakterisiert. Weiterhin wurde der Einfluß von „Ausreißern“ in den Daten unterschätzt und selten wurden Anpassungstests durchgeführt.
Um die aufgetretenen Charakteristika mittels Poissonverteilung zu kompensieren, wurden die Ankunftsprozesse durch die Überlagerungen von vielen unabhängigen Quellen modelliert. Deren Aktivitäten waren mehr oder weniger gedächtnislos. Diese Modellierungen führten zunächst auch zu den gewünschten Ergebnissen, da eine Verbindung im Normalfall aus dem Verbindungsaufbau, der Datenübertragung und der Verbindungsbeendigung bestand.
Seit der Verbreitung des Internets hat sich das Verhalten des Verkehrsaufkommens nochmals verändert. So führt zum Beispiel der Verbindungsaufbau zu einem Webserver nicht nur zu der Übertragung einer einzelnen Datei, sondern zieht den Aufbau weiterer Verbindungen nach sich. Es ist möglich, über mehrere Verbindungen quasi gleichzeitig Daten zu übertragen bzw. gleichzeitig verschiedene Applikationen gestartet zu haben, die auf ein Netzwerk zugreifen. Hier stoßen jedoch Poissonprozesse an ihre Grenze und die Ankunftsprozesse sind nicht weiter mit Poisson zu modellieren.
Tatsächlich zeigt der Ankunftsprozess von TCP-Verbindungen selbstähnliches Verhalten. Die Prämisse der Vorhersagbarkeit war nicht mehr gegeben und der Ansatz der Leitungsvermittlung nicht weiter praktikabel. So gehen neuere Untersuchungen von einem paketorientierten Ansatz aus. In (LTWW93) bzw. (LTWW94) wurde gezeigt, daß die Ankunftsprozesse in Lokalen Netzwerken (kurz: *LAN*) ein selbstähnliches Verhalten aufweisen. Seitdem konnten die Ergebnisse durch weitere Untersuchungen bestätigt und auf Wide Area Netzwerke (Abkürzung: *WAN*) erweitert werden, wie in (PF95),(LWDW97) oder (Fel01) zu sehen.
Der Vorteil der Selbstähnlichkeit liegt in dessen einfachen Beschreibung, mit einem Parameter H. Dieser Parameter ist auch als Hurst-Parameter bekannt. Seither galt es Verteilungen zu finden, die die selbstähnliche Charakteristik berücksichtigen. Als geeignete Verteilungen konnten bisher die Pareto-, Weibull- und Lognormalverteilung identifiziert werden. Hierbei eignet sich die Paretoverteilung vor allem zur Beschreibung der Zwischenankunftszeiten. Aggregierte Ankunftsprozesse sind mit der Lognormalverteilung und Weibullverteilung zu modellieren.
Die Annahme der Selbstähnlichkeit hat sich bereits bei den Simulationswerkzeugen durchgesetzt. So ist OPNET eine Simulationsumgebung mit der unter anderem selbstähnliche Datenströme modelliert werden können. Der innerhalb der Umgebung zur Verfügung stehende Raw Packet Generator (kurz: *RPG*), ist darauf ausgelegt selbstähnlichen Netzverkehr zu simulieren. Dies konnte bereits in einer früheren Arbeit gezeigt werden (vgl. (GK04)).

In keiner betrachteten Veröffentlichung zum Thema der Verteilung von Netzwerkverkehr wird eine Bewertung der Verfahren zur Bestimmung der Verteilungsparameter vorgenommen. Während zumeist der Maximum Likelihood Estimator verwendet wird, gibt es Ausnahmefälle in denen der Least Squares Estimator Verwendung findet, wie in (Den96) zu sehen. Ebenfalls selten sind Anpassungkriterien, um den Grad der Anpassung analytisch zu bestimmen.
In der vorliegenden Arbeit werden diese Aspekte betrachtet. So ergibt sich folgende Gleiderung:

- Einführung einiger grundlegender Begriffe und der betrachteten Verteilungen
- Beschreibung der Verfahren zur Schätzung der Verteilungsparameter
- Vorstellung und Bewertung von verschiedenen Anpassungskriterien
- Darstellung der Verfahren zur Schätzung des Hurst-Parameters
- Beschreibung des Meßsystems und Anwendung der vorgestellten Verfahren

1.2 Grundlagen

In den nächsten Abschnitten werden einige Grundbegriffe genauer erläutert, um das allgemeine Verständnis zu erleichtern. Begonnen wird im folgenden Abschnitt mit dem Begriff der Selbstähnlichkeit.

1.2.1 Selbstähnlichkeit

Für einen einfacheren Einstieg in die Thematik werden in Abbildung 1.1 (aus (WTSW97)) real gemessener Netzwerkverkehr und mittels Poisson modellierter Verkehr vergleichend dargestellt. Gezeigt wird die Anzahl der eingetroffenen Pakete in einem ausgewähltem Intervall. Aus dem Intervall wird jeweils ein beliebiges Subintervall ausgewählt und vergrößert. Wie zu sehen ist, sind sich die Darstellungen des realen Netzwerkverkehrs sehr ähnlich, selbst bei großen Aggregationsstufen. Hier wird von einem selbstähnlichen Verhalten gesprochen, da die „Burstiness“ in allen Zeitbereichen erhalten bleibt. Der poissonmodellierte Ankunftsprozess zeigt kein solches Verhalten, denn hier tritt mit zunehmendem Aggregationsgrad eine Glättung auf.
Um den Begriff der Selbstähnlichkeit (nach (LTWW93)) allgemeiner definieren zu können , sei ein im weitesten Sinne stationärer Prozess $X = (X_t; t = 0, 1, 2, \ldots)$ gegeben, mit konstantem Mittelwert $\mu = E[X_t]$ und einer endlichen Varianz $\sigma^2 = E\left[(X_t - \mu)^2\right]$. Durch

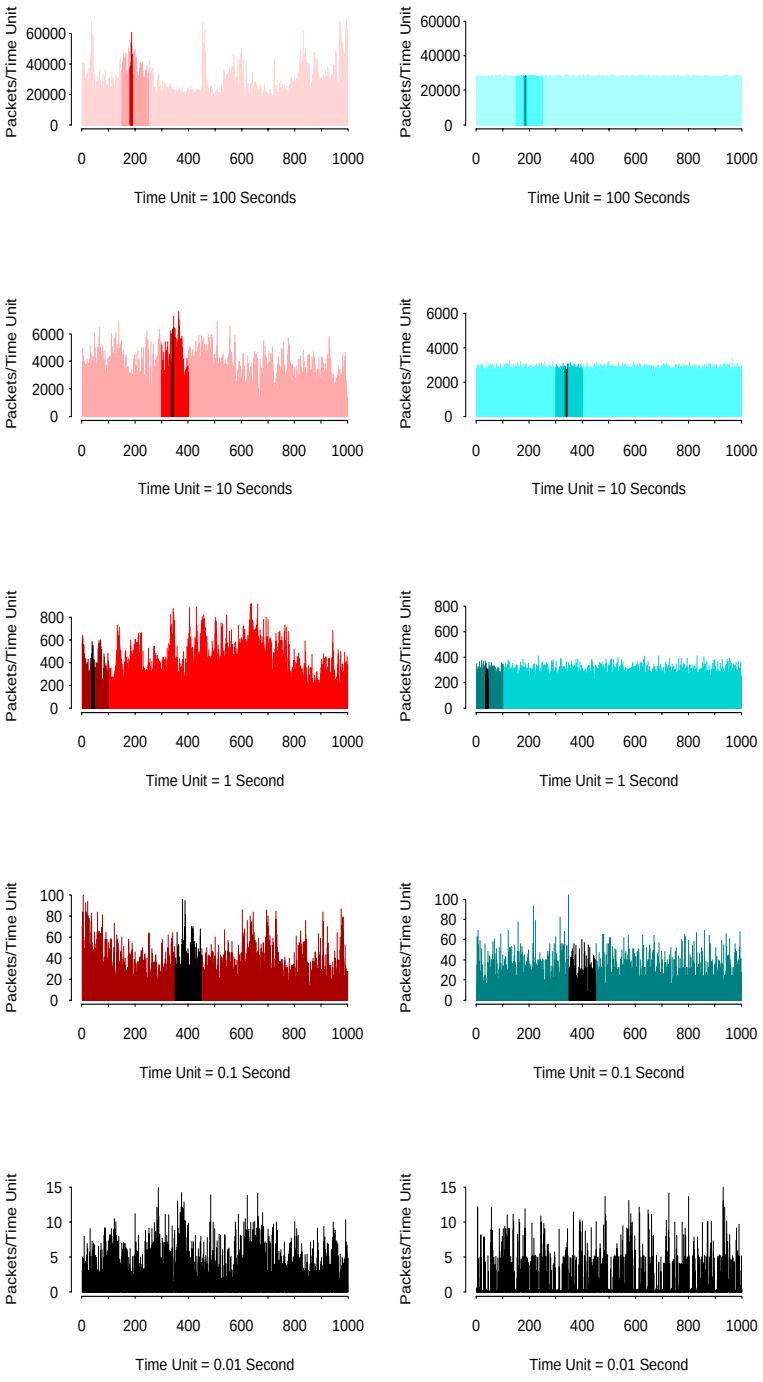


Abbildung 1.1: Gegenüberstellung aus (WTSW97): (links) gemessener Netzwerkverkehr, (rechts) poissonverteilter Netzwerkverkehr

Unterteilung des Prozesses in nicht-überlappende Intervalle (auch: Blöcke) der Größe m und durch Summierung der Werte innerhalb der Blöcke ergibt sich ein m-aggregierter Prozess mit $X^{(m)} = \left(X_k^{(m)} : k = 1, 2, 3, \ldots\right)$. Der hier betrachtete Prozess beschreibt, nach (Fel01) und (LTWW94), die Anzahl der Ankünfte pro Zeiteinheit. Die AKF von X ist definiert als

$$r(k) = E\left[\left(X_t - \mu\right)\left(X_{t+k} - \mu\right)\right]. \tag{1.1}$$

Im allgemeinen wird zwischen exakter und asymptotischer Selbstähnlichkeit unterschieden. Ein Prozess X ist exakt selbstähnlich, wenn die zugehörigen aggregierten Prozesse $X^{(m)}$ die gleiche Korrelationsstruktur wie X besitzen. Es muß gelten:

$$r^{(m)}(k) = r(k) \text{ für alle } m = 1, 2, 3, \ldots \text{ und } k = 1, 2, 3, \ldots \tag{1.2}$$

Somit ist X exakt selbstähnlich, wenn die aggregierten Prozesse $X^{(m)}$ nicht von X zu unterscheiden sind. Die Autokorrelationsfunktion ist somit nur abhängig von k. Als Beispiel für exakte Selbstähnlichkeit sei hier Fractional Gaussian Noise (FGN) zu erwähnen mit $0.5 < H < 1$. H entspricht dem Grad der Selbstähnlichkeit (vgl. Abschnitt 1.2.3).
Asymptotische Selbstähnlichkeit für einen stationären Prozess X ist gegeben, wenn $r^{(m)}$ asymptotisch übereinstimmt. Für große k und $m \to \infty$ wird die Korrelationsbedingung aus Gleichung [1.2] abgeschwächt zu:

$$r^{(m)}(k) \to r(k) \tag{1.3}$$

FARIMA(p,d,q)-Prozesse mit $0 < d < \frac{1}{2}$ sind Beispiele für asypmtotisch selbstähnliche Prozesse.
Unter anderem in (Fel01), (PF95) und (LTWW94) wurde asymptotische Selbstähnlichkeit im Netzwerkverkehr (LAN und WAN) nachgewiesen. In (LTWW94) werden selbstähnliche Prozesse durch Überlagerung von mehreren ON/OFF-Quellen beschrieben. Die Verteilung der ON- bzw. OFF-Perioden muß dann jedoch „heavy-tailed“ sein. Auch die Simulationsumgebung OPNET verwendet dieses Modell (vgl. (Ryu00)).

1.2.2 Langzeitabhängigkeit (LRD)

Der Begriff der Langzeitabhängigkeit (engl.: *long range dependence*) ist eng mit der Selbstähnlichkeit verknüpft und ist auch unter den Bezeichungen *Joseph-Effect* oder *Heavy-Tailedness* bekannt. Die Langzeitabhängigkeit in paketvermittelten Datennetzen zeichnet sich durch starke Schwankungen der Ankunftszeiten aus, die nicht mit Hilfe der Exponentialverteilung erfasst werden können.
In diesem Zusammenhang tritt in vielen Veröffentlichungen auch der Begriff „*burstiness*“ auf. Dieser beschreibt die Eigenschaft, daß Phasen mit niedrigem Verkehrsaufkommen sporadisch abgelöst werden von Phasen mit hohem Verkehrsaufkommen.
Ist ein Prozess statistisch selbstähnlich, so zeigt sich auch LRD (vgl. (LTWW93)). Wie bereits in Abschnitt 1.2.1 gezeigt, ist die Autokorrelationsfunktion einzig abhängig von k. Somit kann die Autokorrelationsfunktion eines Prozesses mit Langzeitabhängigkeit wie folgt beschrieben werden:

$$r(k) = c_1 \cdot k^{-\beta} \text{ mit } k \to \infty \text{ und } 0 < \beta < 1.$$

Hier ist c_1 als Konstante anzusehen. Folglich nimmt die AKF eines solchen Prozesses polynomial ab, wie in Abbildung 1.2 aus (LWDW97) zu sehen. Für Prozesse mit SRD (*short range dependence*) ist ein exponentieller Rückgang zu verzeichnen. Dabei darf die

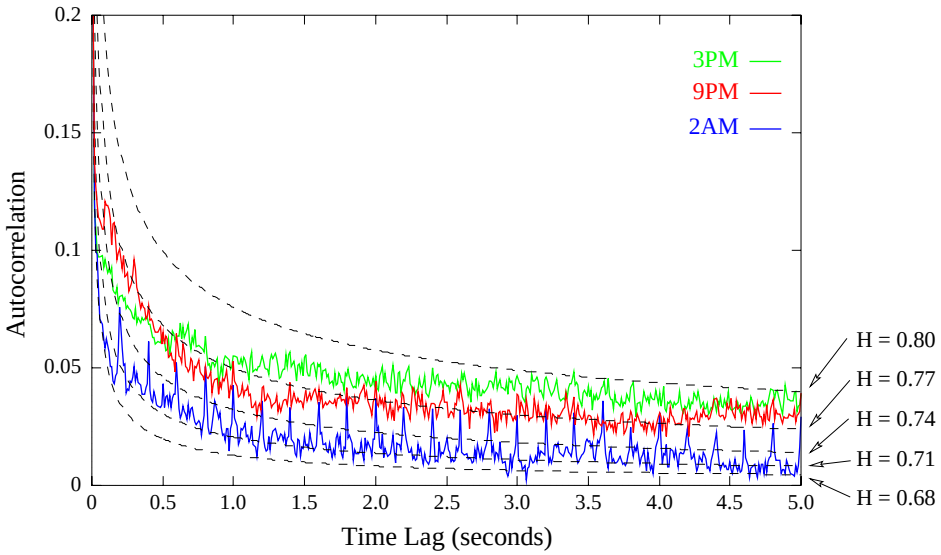


Abbildung 1.2: Darstellung von Autokorrelationsfunktionen aus (LWDW97)

Varianz ebenfalls nur polynomial abnehmen. Dieser Zusammenhang wird beim Variance-Time-Plot ausgenutzt (vgl. Abschnitt 4.2). Unter anderem in (LTWW94) und (Fel01) wurde gezeigt, daß Netzwerkverkehr Langzeitabhängigkeit aufweist. In (CB95) werden für das Auftreten von Langzeitabhängigkeit folgende Gründe angegeben:

- Verteilung der Dateigrößen
- Einflüße durch caching
- Verhalten der Benutzer bei Datentransfer
- Bedenkzeiten der Benutzer

Dabei wurden jedoch nur HTTP-Verbindungen berücksichtigt. Andere Anwendungen wie *Video-on-Demand* oder *Streaming Audio* erzeugen ebenfalls eine zeitlich schwankende Last.

1.2.3 Hurst-Parameter

In einer Vielzahl von Veröffentlichungen wird der Hurst-Parameter ausführlich beschrieben, unter anderem in (CB95), (LTWW94) und (KFR02).
Die Verwendung von selbstähnlichen Modellen zur Beschreibung von Zeitreihen hat den Vorteil, daß der Grad der Selbstähnlichkeit mit nur einem Parameter H ausgedrückt werden kann. Der als Hurst-Parameter bekannte Wert gibt an, wie schnell die Autokorrelationsfunktion der Zeitreihe abnimmt, in Abhängigkeit vom Grad der Aggregation (vgl. Abschnitt 1.2.1 und 1.2.2).
Für selbstähnliche Zeitreihen gilt $0,5 < H < 1$, wobei der Grad der Selbstähnlichkeit mit $H \to 1$ ansteigt. Weicht der Hurst-Parameter signifikant von 0,5 ab, kann von Langzeitabhängigkeiten und somit von Selbstähnlichkeit ausgegangen werden. Folgende Verfahren zur Schätzung des Hurst-Parameters werden in Kapitel 4 näher vorgestellt:

- R/S-Statistik
- Variance-Time-Plot
- Periodogramm

Weitere Methoden existieren, werden jedoch in Zusammenhang mit dieser Arbeit nicht betrachtet.

1.2.4 Stationarität

Die meisten statistischen Verfahren gehen von der Stationarität der zu betrachtenden Meßreihe aus. Eine Zeitreihe $x(t)$ mit der Wahrscheinlichkeitsverteilung $p(x)$ heißt stark stationär, wenn für alle k die k-ten bedingten Wahrscheinlichkeiten nicht von der Zeit abhängen. Es gilt:

$$p\left(x_{t_1+k}|x_{t_1+k-1},\ldots,x_k\right) = p\left(x_{t_2+k}|x_{t_2+k-1},\ldots,x_k\right) \text{ mit } t_1, t_2 \text{ beliebig.}$$

Schlußfolgernd ist für starke Stationarität abzuleiten, daß für alle Momente M_k

$$M_k = \int_{-\infty}^{+\infty} x(t)^k \cdot p(x)dx \neq M_k(t) \tag{1.4}$$

gilt. Sie sind somit unabhängig von der Zeit. Momente entsprechen hierbei Kenngrößen in der Statistik, mit denen Verteilungsfunktionen beschrieben werden können. Folgende Momente sind definiert:

- Erstes Moment: Erwartungswert

- Zweites Moment: Varianz
- Drittes Moment: Schiefe
- Viertes Moment: Wölbung, auch: Kortosis, auch: Exzess

Wölbung und Schiefe sind als Maß der Abweichung einer Verteilung von der Normalverteilung zu sehen, da die Normalverteilung allein durch den Erwartungswert und durch die Varianz bestimmt ist.
Weiterhin muß für starke Stationarität gelten:

$$\phi(t_1, t_2) = \int_{-\infty}^{+\infty} x(t + t_1) \cdot x(t + t_2)dt = \phi(t_1 - t_2). \tag{1.5}$$

Die Autokovarianz hängt nur von der Differenz ihrer Argumente ab. Als Beispiele für stark stationäre Prozesse sind weißes Rauschen und ARMA-Modelle zu nennen.
Im allgemeinen ist die Überprüfung auf starke Stationarität in experimentellen Daten nicht möglich, da unendlich viele Tests erforderlich wären. Weiterhin kann bei einer Stichprobe nicht davon ausgegangen werden, daß alle real auftretenden Werte auch in der Stichprobe enthalten sind. Somit ist die Stationarität in einem weiteren Sinn zu betrachten.
Eine Zeitreihe heißt schwach stationär, wenn ihr erstes und zweites Moment nicht von der Zeit abhängen. Somit muß gelten:

$$\mu = konstant \text{ und } \sigma^2 = konstant$$

Ein Verfahren zur Untersuchung auf Stationarität ist die Fensterbildung. In verschiedenen Zeitbereichen gleicher Länge werden die Mittelwerte und Varianzen untersucht. Unter anderem in (Fel01) wurde diese Methode angewandt. Auch in Zusammenhang mit dieser Arbeit wird diese Methode in Abschnitt 5.4 betrachtet.

1.3 Betrachtete Verteilungen

Die Verteilungsfunktion einer Zufallsvariablen ist nach (Pax94) definiert mit

$$F(x) = P\left(X \leq x\right). \tag{1.6}$$

$F(x)$ ist die Wahrscheinlichkeit, daß eine Merkmalsausprägung (bzw. Instanz) der Zufallsvariablen X einen Wert kleiner oder gleich x annimmt.
Bis 1995 wurden die Ankunftszeiten fast ausschließlich mittels Exponentialverteilung modelliert. Handelte es sich dementsprechend um aggregierte Ankunftsprozesse, wurde die Poissonverteilung angewandt. In den letzten Jahren hat sich jedoch gezeigt, daß diese Annahme unzureichend ist, da die Bedingung der strikten Gedächtnislosigkeit zumeist nicht gegeben ist. In (Fel01) wird gezeigt, daß „heavy-tailed"-Verteilungen hier eine bessere Anpassung ermöglichen.
Zu den Verteilungen, denen die Eigenschaft der „heavy-tailedness" zugeschrieben werden kann, gehören die:

- Paretoverteilung
- Weibullverteilung
- Lognormalverteilung[1]

In den nächsten Abschnitten werden die oben aufgelisteten Verteilungen vorgestellt. Vergleichend hierzu wird ebenfalls genauer auf die Exponentialverteilung als einzige einparametrige Verteilung eingegangen. Aufgezeigt werden die Parameter und deren Einfluß auf die Form und Lage der Verteilungen. Die Methoden zur Bestimmung der Parameter werden in Kapitel 2 ausführlich dargestellt.

1.3.1 Die Exponentialverteilung

In der Warteschlangentheorie kommt der Exponentialverteilung eine besondere Bedeutung zu. Bei der leitungsvermittelten Telefonie sind die Gesprächsdauer und die Anzahl der Leitungen in einer Vermittlungsstelle in Abhängigkeit von der Zeit exponentialverteilt. Ausgegangen wird von der Gedächtnislosigkeit des Kanals und somit der Verteilung. Ereignisse in der Vergangenheit haben keinen Einfluß auf Ereignisse in der Zukunft. Daraus folgend nimmt die Autokorrelationsfunktion des Prozesses exponentiell ab. Unter anderem in (LTWW93) wurde gezeigt, daß in paketvermittelten Netzen die Voraussetzung

1 umstritten - siehe Abschnitt 1.3.2

der Gedächtnislosigkeit nicht gegeben ist. Versuche, paketvermittelten Verkehr mit Hilfe der Exponentialverteilung zu modellieren, konnten keine realistischen Resultate über alle Zeitbereiche erzielen, da es sich hierbei um selbstähnliche Ankunftsprozesse handelte. Die Dichtefunktion der Exponentialverteilung ist nach (BSMM99) gegeben mit

$$f(x) = \lambda \cdot e^{-\lambda x} \text{ für } \lambda > 0,\ x \geq 0 \tag{1.7}$$

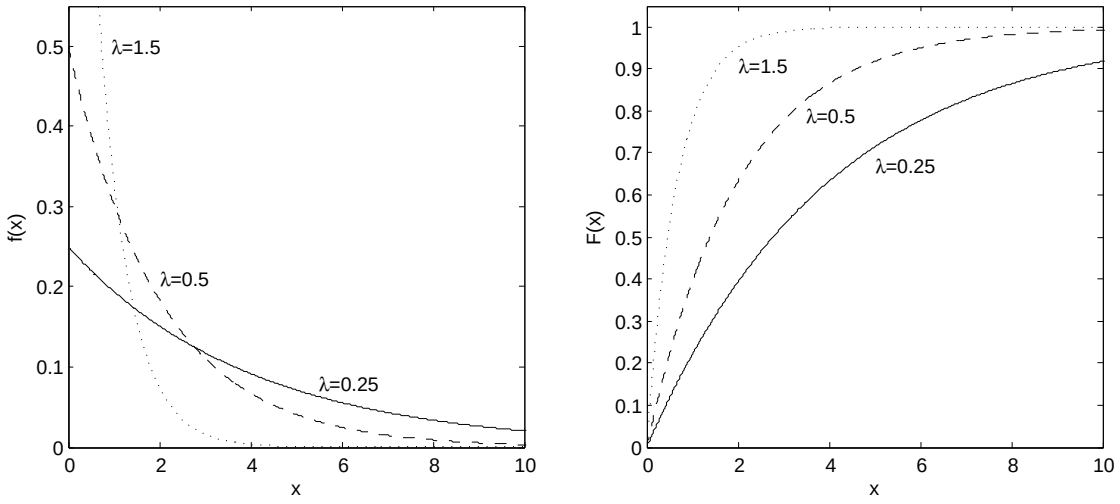
und die Wahrscheinlichkeitsverteilung mit

$$F(x) = 1 - e^{-\lambda x} \tag{1.8}$$

Im allgemeinen wird x als Zeit angenommen. Der Parameter λ ist gegeben durch $\lambda = 1/\mu$ mit μ als Intensität (z.B. mittlere Ankunftsrate). In Abbildung 1.3 sind die Dichtefunktion und Verteilungsfunktion für verschiedene Werte für λ dargestellt. Das einzige für diese Arbeit relevante Moment ist der Erwartungswert und ist definiert als:

$$EX = \mu = \frac{1}{\lambda}. \tag{1.9}$$

Die Aggregation einer exponentialverteilten Zufallsvariable resultiert in der Poissonverteilung.



(a) *Dichtefunktion* **(b) *Verteilungsfunktion***

Abbildung 1.3: Einfluß des Parameters der Exponentialverteilung

1.3.2 Die logarithmische Normalverteilung

Die logarithmische Normalverteilung (auch: *Lognormalverteilung*) besitzt nach (CB95) nicht die Eigenschaft der unendlichen Varianz, womit die Lognormalverteilung nicht zu den „heavy-tailed“ Verteilungen zu zählen ist. Über diese Aussage existiert jedoch keine übereintimmende Meinung, denn in (LWDW97) und (Fel01) wird der Lognormalverteilung die Eigenschaft der „heavy-tailedness“ zugeschrieben.
Die Lognormalverteilung ist eng mit der Normalverteilung verknüpft. Ist der Logarithmus einer Zufallsvariablen X normalverteilt, so ist X lognormalverteilt. Nach (BSMM99) und (Sch01) ist die Dichtefunktion gegeben mit:

$$f(x) = \frac{\log(e)}{x\sigma_L\sqrt{2\pi}} e^{-\frac{(\log(x)-\mu_L)^2}{2{\sigma_L}^2}} \text{ für } x > 0.$$

In der Praxis wird entweder der natürliche oder der dekadische Logarithmus verwendet. Die hier folgenden Betrachtungen beziehen sich auf den natürlichen Logarithmus. So ist die Dichtefunktion gegeben mit:

$$f(x) = \frac{1}{x\sigma_L\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu_L)^2}{2{\sigma_L}^2}} \text{ für } x > 0. \tag{1.10}$$

Während bei der Normalverteilung μ ein Lageparameter und σ ein Skalierungsparameter ist, entspricht bei der Lognormalverteilung der Parameter μ_L einem Skalierungsparameter und σ_L einem Formparameter (vgl. Abb. 1.4). Die Verteilungsfunktion ist definitiert als:

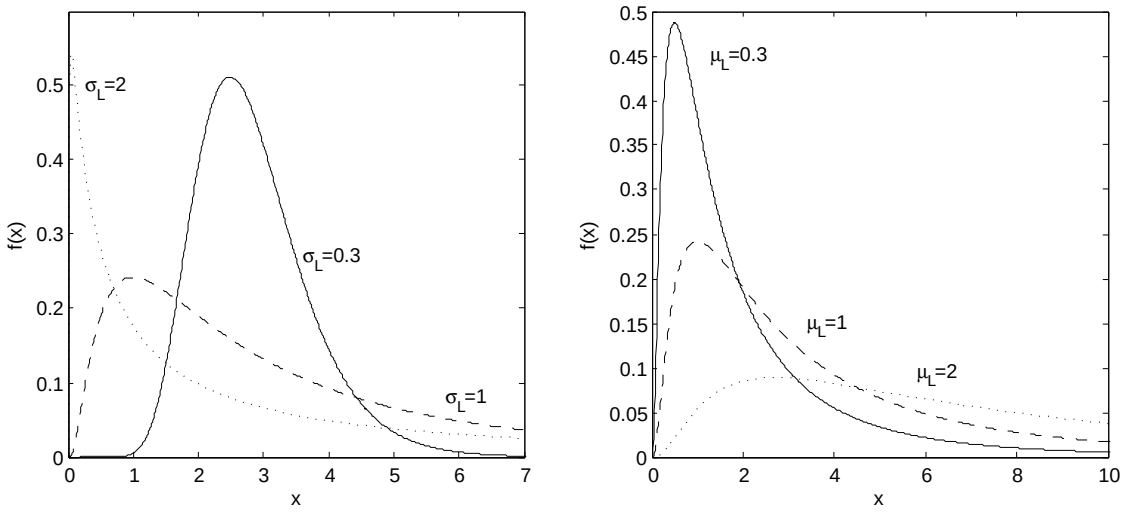
$$F(x) = \frac{1}{\sigma_L\sqrt{2\pi}} \int_{-\infty}^{\ln x} e^{-\frac{(t-\mu_L)^2}{2{\sigma_L}^2}} dt. \tag{1.11}$$

Für die Schätzung der Parameter mit Hilfe der Methode der Momente (siehe Abschnitt 2.3) werden der Erwartungswert und die Streuung benötigt, für die nach (BSMM99) gilt:

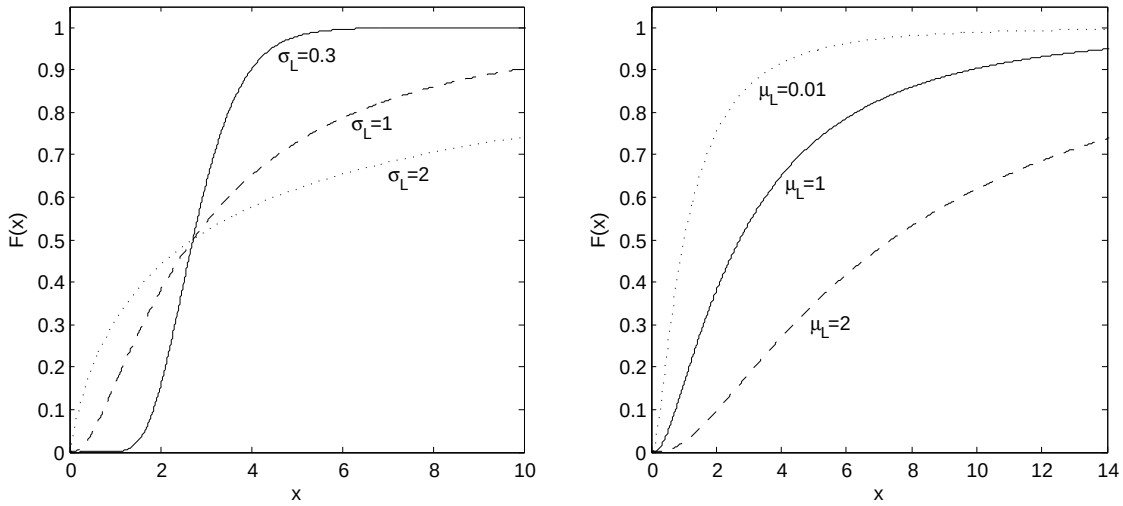
$$EX = \mu = e^{\left(\mu_L + \frac{{\sigma_L}^2}{2}\right)} \tag{1.12}$$

$$Var\ X = \sigma^2 = e^{\left(2\mu_L + {\sigma_L}^2\right)} \cdot \left(e^{{\sigma_L}^2} - 1\right). \tag{1.13}$$

Hierbei sind μ_L und σ_L die Parameter der Lognormalverteilung, während μ und σ die Parameter der Normalverteilung sind. In Abbildung 1.4 werden die Dichte- und Verteilungsfunktionen der Lognormalverteilung dargestellt, mit verschiedenen Werten für μ_L und σ_L. In (LWDW97) wird gezeigt, daß die dort untersuchten, aggregierten Zwischenankunftszeiten in einem WAN lognormalverteilt sind.



(a) *Dichtefunktion,* $\mu_L = 1$ **(b) *Dichtefunktion,*** $\sigma_L = 1$



(c) *Verteilungsfunktion,* $\mu_L = 1$ **(d) *Verteilungsfunktion,*** $\sigma_L = 1$

Abbildung 1.4: Einfluß der Parameter der Lognormalverteilung

1.3.3 Die Paretoverteilung

Aufgrund ihrer selbstähnlichen Charakteristik (vgl. (PF95)) wurde die Paretoverteilung, benannt nach Vilfredo Pareto, in vielen Veröffentlichungen gewählt, um langzeitabhängige Prozesse zu erzeugen. Die Paretoverteilung ist die einfachste Verteilung mit „heavy-tailed“ Eigenschaften. Die Dichtefunktion ist gegeben mit

$$f(x) = \frac{\alpha_p}{t_0} \cdot \left(\frac{t_0}{x}\right)^{\alpha_p+1} \text{ für } x > t_0,\ t_0 > 0. \tag{1.14}$$

Hierbei ist α_p der Formparameter und t_0 der Lageparameter. Die Verteilungsfunktion der Paretoverteilung ist definiert durch:

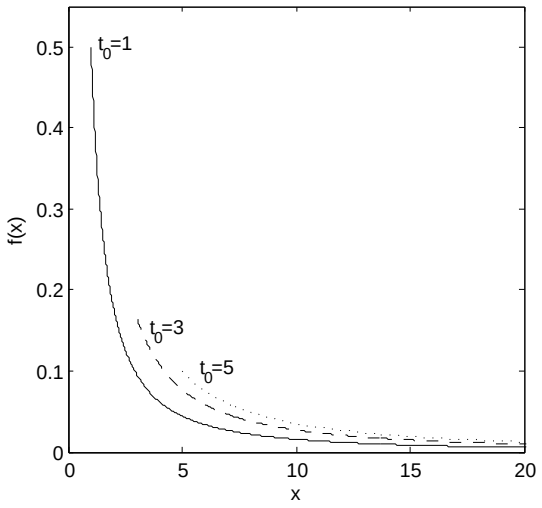
$$F(x) = 1 - \left(\frac{t_0}{x}\right)^{\alpha_p} \text{ für } x > t_0,\ t_0 > 0 \tag{1.15}$$

Für die Methode der Momente werden wiederum die ersten beiden Momente benötigt, die wie folgt definiert sind:

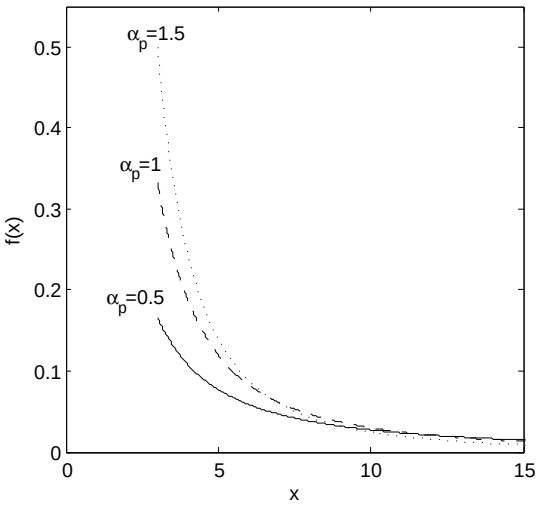
$$EX = \mu = \frac{\alpha_p t_0}{\alpha_p - 1} \text{ mit } \alpha_p > 1 \tag{1.16}$$

$$Var\ X = \sigma^2 = \frac{\alpha_p {t_0}^2}{(\alpha_p - 2)\,(\alpha_p - 1)^2}. \tag{1.17}$$

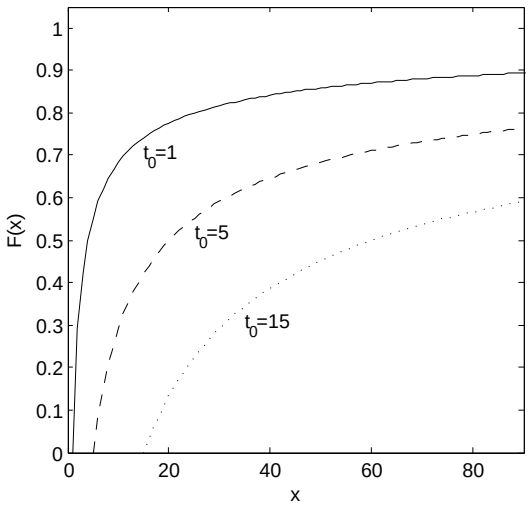
EX entspricht dem Erwartungswert und $Var\ X$ der Varianz der Verteilung. Der Einfluß der Parameter auf die Verteilungsdichtefunktion und Verteilungsfunktion, ist in Abbildung 1.5 zu erkennen. Dabei wurde jeweils ein Parameter mit einem konstanten Wert versehen und der zweite Parameter dementsprechend variiert.
Der Paretoverteilung kommt in OPNET eine besondere Bedeutung zu. Der RPG basiert auf der Paretoverteilung, dessen Parameter t_0 als Parameter FOTS (*Fractal Onset Time Scale*) im RPG verwendet wird.



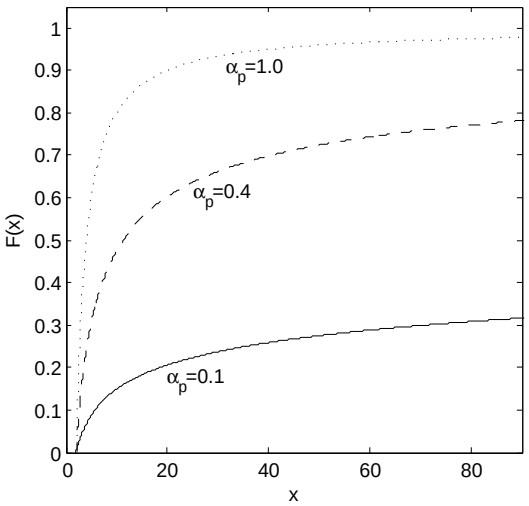
(a) *Dichtefunktion,* $\alpha_p = 0,5$



(b) *Dichtefunktion,* $t_0 = 3$



(c) *Verteilungsfunktion,* $\alpha_p = 0,5$



(d) *Verteilungsfunktion,* $t_0 = 2$

Abbildung 1.5: Einfluß der Parameter der Paretoverteilung

1.3.4 Die Weibullverteilung

Die Weibullverteilung wurde anfangs von dem schwedischen Physiker Waloddi Weibull verwendet, um die Bruchfestigkeit von verschiedenen Materialien zu beschreiben. Seitdem wurden auch andere Anwendungsgebiete, wie die Beschreibung von Windschwankungen, erschlossen. Durch entsprechende Anpassung der Parameter konnte in (Fel01) der Nachweis erbracht werden, daß auch die Ankunftszeiten von TCP-Netzwerkverkehr mittels Weibullverteilung beschrieben werden können. Andere dort betrachtete Verteilungen (Pareto und Lognormal) wichen zum Teil sehr deutlich ab. Unter anderem in (JKB70a), (BSMM99) und (Fel01) wird die Verteilungsdichtefunktion angegeben mit

$$f(x) = \frac{\alpha_w}{\beta} \cdot \left(\frac{x}{\beta}\right)^{\alpha_w - 1} \cdot exp\left(-\left(\frac{x}{\beta}\right)^{\alpha_w}\right) \text{ für } \alpha_w > 0,\ \beta > 0,\ x \geq 0. \tag{1.18}$$

Die Wahrscheinlichkeitsverteilung ist somit definiert als

$$F(x) = 1 - e^{-(\frac{x}{\beta})^{\alpha_w}}. \tag{1.19}$$

Hier ist α_w der Formparameter und β der Vergrößerungsparameter. Der Formparameter ist nicht mit dem Parameter α_p der Paretoverteilung gleichzusetzen. Die Abhängigkeit der Dichtefunktion und Verteilungsfunktion von den Parametern ist der Abbildung 1.6 zu entnehmen.

Für die Methode der Momente sind die ersten beiden Momente der Weibullverteilung anzugeben. Der Erwartungswert und die Varianz sind gegeben mit

$$EX = \mu = \beta \cdot \Gamma\left(1 + \frac{1}{\alpha_w}\right) \tag{1.20}$$

und

$$Var\ X = \sigma^2 = \beta^2 \left(\Gamma\left(1 + \frac{2}{\alpha_w}\right) - \Gamma^2\left(1 + \frac{1}{\alpha_w}\right)\right). \tag{1.21}$$

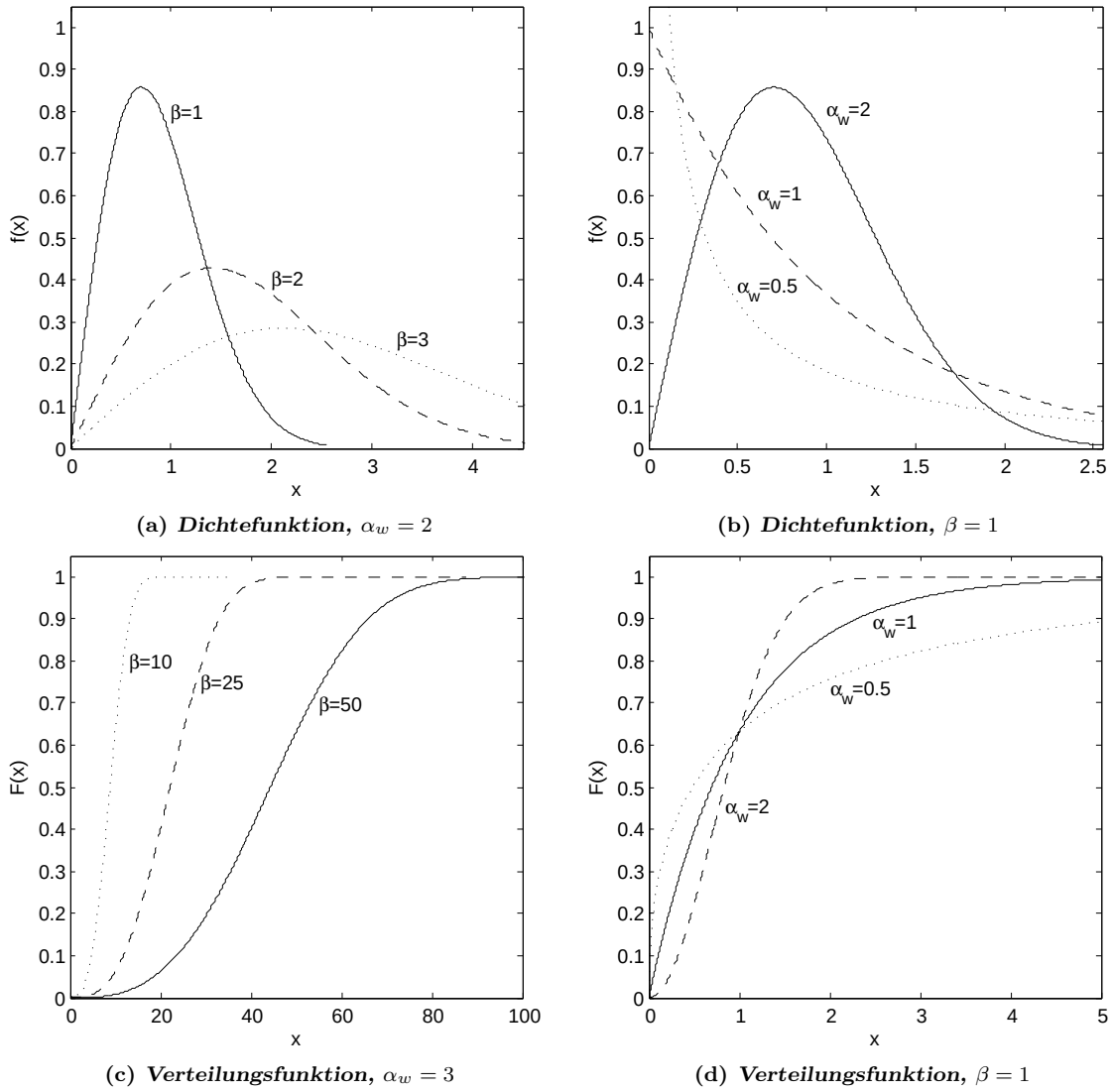
Γ beschreibt die Gammafunktion, für die gilt:

$$\Gamma(x) = \int_{x=0}^{\infty} t^{x-1} e^{-t} dt \text{ für } x > 0. \tag{1.22}$$

Die Weibullverteilung existiert ebenfalls als 3-parametrige Form, auf die in dieser Arbeit nicht eingegangen wird. Eingefügt wird der Parameter γ, als Lageparameter.

Aus der Dichtefunktion und der Verteilungsfunktion wird ersichtlich, daß es sich bei der Exponentialverteilung um eine spezielle Form der Weibullverteilung handelt. Mit $\alpha_w = 1$ und $\beta = \mu$ ergibt sich hier die Exponentialverteilung.

Für $\alpha_w = 2$ und $\beta = 1/(2b^2)$ zeigt sich nach (HEK93) ein weiterer Spezialfall der Weibullverteilung, die Rayleighverteilung. Auf diese Verteilung mit dem Parameter b soll im Zusammenhang dieser Arbeit nicht genauer eingegangen werden.



(a) *Dichtefunktion,* $\alpha_w = 2$

(b) *Dichtefunktion,* $\beta = 1$

(c) *Verteilungsfunktion,* $\alpha_w = 3$

(d) *Verteilungsfunktion,* $\beta = 1$

Abbildung 1.6: Einfluß der Parameter der Weibullverteilung

1.3.5 Übersicht der Dichte- und Verteilungsfunktionen

In der folgenden Tabelle 1.1 werden die Dichte- und Verteilungsfunktionen der einzelnen Verteilungen vergleichend gegenübergestellt. Eine Zusammenfassung der Momente

Verteilung	Dichtefunktion	Verteilungsfunktion
Exponential	$f(x) = \lambda e^{-\lambda x}$	$F(x) = 1 - e^{-\lambda x}$
Lognormal	$f(x) = \frac{1}{x\sigma_L\sqrt{2\pi}} exp\left(-\frac{(\ln(x)-\mu_L)^2}{2{\sigma_L}^2}\right)$	$F(x) = \frac{1}{\sigma_L\sqrt{2\pi}} \int_{-\infty}^{\ln x} exp\left(-\frac{(t-\mu_L)^2}{2{\sigma_L}^2}\right)dt$
Pareto	$f(x) = \frac{\alpha_p}{t_0} \cdot \left(\frac{t_0}{x}\right)^{\alpha_p+1}$	$F(x) = 1 - \left(\frac{t_0}{x}\right)^{\alpha_p}$
Weibull	$f(x) = \frac{\alpha_w}{\beta} \cdot \left(\frac{x}{\beta}\right)^{\alpha_w-1} \cdot exp\left(-\left(\frac{x}{\beta}\right)^{\alpha_w}\right)$	$F(x) = 1 - exp\left(-\left(\frac{x}{\beta}\right)^{\alpha_w}\right)$

Tabelle 1.1: Übersicht der Dichte- und Verteilungsfunktionen

der Verteilungen ist in Abschnitt 2.3 zu finden, da diese für die Schätzung der Parameter relevant sind. Im nächsten Kapitel werden die Schätzverfahren zur Bestimmung der Parameter detailiert vorgestellt.

Kapitel 2

Verfahren zur Bestimmung der Verteilungsparameter

Ist eine geeignete Verteilung ausgewählt, sind die Parameter zu schätzen. Eine Vorauswahl kann bereits mit der Methode aus Abschnitt 2.1 erfolgen. In den nächsten Abschnitten werden die am häufigsten verwendeten Verfahren zur Schätzung der Parameter vorgestellt. Dabei wird auf deren Vor- und Nachteile eingegangen. Zu den hier untersuchten Verfahren zählen:

- der Least Squares Estimator - LSE
- die Methode der Momente
- der Maximum Likelihood Estimator - MLE

Betrachtet werden diese Verfahren jeweils für die Exponential-, Lognormal-, Pareto- und Lognormalverteilung. In einigen Fällen ist es sinnvoll, die Parameter der Verteilungen nur für bestimmte, dominierende Bereiche zu betrachten, wie in (Pax94) beschrieben. Vor allem bei der Pareto-Verteilung wird dies nötig sein, da sie in dem Bereich $x < t_0$ nicht definiert ist (vgl. 1.3.3).

2.1 Probability Plotting

Beim Probability Plotting (auch: *Methode des Wahrscheinlichkeitspapiers*) handelt es sich um eine graphische Methode zur Bestimmung der Verteilungsparameter. Konstruiert wird eine spezielle Darstellung der CDF (Cumulative Distribution Function) aus den Daten. Durch eine geeignete Achsentransformation ist die CDF in eine Gerade der Form $y = ax + b$ zu überführen (vgl. (HEK93)). Wie in Abschnitt 2.2.3 gezeigt wird, gilt für die 2-parametrige Weibullverteilung

$$F_x(x) = 1 - e^{-(\frac{x}{\beta})^{\alpha_w}}$$

nach linearer Transformation

$$\ln(-\ln(1 - F_x(x))) = \alpha_w \cdot \ln(x) - \alpha_w \cdot ln(\beta).$$

Aus der linearisierten Form der Verteilungsfuntkion sind transformierte Achsen zu bestimmen. Am Beispiel der Weibullverteilung ergeben sich die Parameter zu:

$$\begin{aligned} y &= \ln(-\ln(1 - F_x(x))) \\ x &= \ln(x) \\ a &= \alpha_w \\ b &= -\alpha_w \cdot \ln(\beta). \end{aligned}$$

Die hier gezeigten Transformationen der Abszissen- und Ordinatenachse werden auf einen Graphen angewandt. Anschließend sind die Punkte zu setzen und eine Gerade zu approximieren. Durch Ablesen des Anstiegs und des Schnittpunktes der Geraden mit der Ordinatenachse ergibt sich die Geradengleichung. Die Parameter der Verteilung sind aus den Parametern der Geraden zu bestimmen. Am Beispiel der Weibullverteilung ergibt sich für die Verteilungsparameter:

$$\hat{\alpha}_w = a \text{ und } \hat{\beta} = e^{-\frac{b}{\hat{\alpha}_w}}.$$

Es ist zu erkennen, daß es sich hierbei um keine genaue Methode handeln kann. Mit Hilfe dieser Methode können nur Anhaltspunkte für die wahren Werte der Parameter gewonnen werden. Es ist mit weiteren Verfahren zu prüfen, inwiefern die Schätzungen eine genaue Anpassung ermöglichen. Der Anstieg der Geraden ließe sich auch durch lineare Regression bestimmen; dieses ist die Vorgehensweise des Least Squares Estimators, wie Abschnitt 2.2 gezeigt wird.
Der Vorteil dieser Methode liegt in der graphischen Darstellung. Hier kann mit geringem mathematischen Aufwand im Vorfeld eine Prognose erfolgen, ob eine gewählte Verteilung die Daten hinreichend genau beschreibt. Zeichnen die dargestellten Punkte näherungsweise eine Gerade, kann die Verteilung als mögliche Verteilung angenommen werden. Ist ein

Bogen zu erkennen, ist die gewählte Verteilung in den meisten Fällen nicht zutreffend. Verallgemeinert werden kann diese Schlußfolgerung jedoch nicht.
Aufgrund der Tatsache, daß die Verteilungsfunktion der Lognormalverteilung (vgl. Gleichung [1.11]) nicht in geschlossener Form vorliegt, ist es nicht möglich eine geeignete lineare Transformation durchzuführen. So wird die Methode des Probability Plottings nur im Zusammenhang mit der Lognormalverteilung verwendet. Aufgrund des Zusammenhangs zwischen Normal- und Lognormalverteilung (vgl. Abschnitt 1.3.2) kann der Probability Plot der Normalverteilung angewandt werden. In MATLAB steht die Funktion *normplot* zur Verfügung.
Die weiteren Verteilungen werden mit dem LSE-Verfahren im nächsten Abschnitt behandelt.

2.2 Methode der kleinsten Quadrate

Die Methode der kleinsten Quadrate (auch: *„Least Squares Estimator“*, kurz: *LSE*) greift die Idee des Probability Plottings auf. Wie jedoch in Abschnitt 2.1 gezeigt, werden die Kennwerte der Geraden aus der Darstellung entnommen. In diesem Zusammenhang ist die Methode der kleinsten Quadrate als eine Erweiterung der obigen Methode zu sehen, da hier eine lineare Regression zur Bestimmung des Anstiegs durchgeführt wird.
Einer Menge von Punkten $(x_1, y_1)..(x_n, y_n)$ wird wiederum eine Gerade zugeordnet. Die Minimierung der Summe der Quadrate zwischen den Punkten der Darstellung und der angenommenen Gerade kann in horziontaler und vertikaler Richtung erfolgen. Abbildung 2.1 soll dies verdeutlichen. Gesucht ist wiederum eine Gerade der Form $y = \hat{a}x + \hat{b}$ die

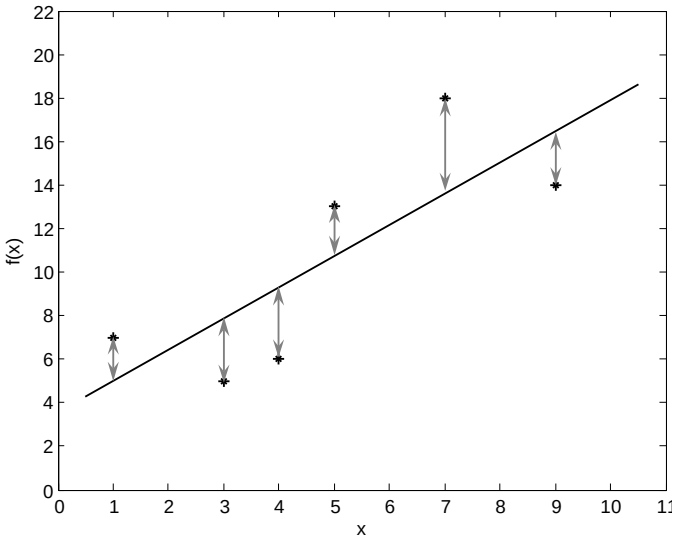


Abbildung 2.1: Minimierung der vertikalen Abstände

folgende Bedingung erfüllt:

$$\sum_{i=1}^{N} [y_i - (b + ax_i)]^2 = min. \tag{2.1}$$

Hierbei sind $\hat{a}$, $\hat{b}$ die geschätzten Werte der Parameter der Geraden (vgl. (BSMM99)). Weiterhin gibt N die Anzahl der Punkte an. Die Parameter der Geraden a, c ergeben sich zu

$$\hat{a} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \quad \text{und} \quad \hat{b} = \bar{y} - \hat{a}\bar{x} \tag{2.2}$$

mit

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{und} \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{2.3}$$

Die Bestimmung der Parameter der Verteilung wird detailiert in den folgenden Abschnitten vorgestellt. In (Pax94) wird diese Methode verwendet, um die Parameter der Paretoverteilung zu schätzen. Der Nachteil dieser Methode liegt in der Definition des Regressionsbereiches. Eine Regression über den Gesamtbereich kann zu ungenauen Werten führen. Eine Einschränkung des Regressionsbereich führt dazu, daß die resultierenden Parameter subjektiv beeinflußt werden. Im allgemeinen ist der LSE als ungenaues Mittel zur Bestimmung der Parameter zu sehen.

2.2.1 Anwendung für die Exponentialverteilung

In diesem Abschnitt wird die Herleitung des LSE für die Exponentialverteilung dargestellt. Mit der Verteilungsfunktion der Exponentialverteilung (vgl. Gleichung [1.8]) ergibt sich durch lineare Transformation:

$$\begin{aligned} (1 - F_x(x)) &= e^{-\lambda x} \\ ln(1 - F_x(x)) &= -\lambda x \\ -ln(1 - F_x(x)) &= \lambda x \end{aligned}$$

Die Parameter der linearisierten Verteilungsfunktion sind zu berechnen nach:

$$\begin{aligned} y &= -ln(1 - F_x(x)) \\ x &= x \\ a &= \lambda \\ b &= 0 \end{aligned}$$

Das Ergebnis der Herleitung stimmt mit (Hab02) überein. Da es sich bei der Expontentialverteilung um eine einparametrige Wahrscheinlichkeitsverteilung handelt, entspricht der Parameter λ dem Anstieg der Geraden

$$\hat{\lambda} = a. \tag{2.4}$$

2.2.2 Anwendung für die Paretoverteilung

Zur Bestimmung der Parameter der Paretoverteilung wird in (JKB70a) unter anderem die Methode der kleinsten Quadrate vorgeschlagen. Ausgehend von der Verteilungsfunktion der Paretoverteilung (vgl. [1.15]) ergibt sich die linearisierte Form der Paretoverteilung nach:

$$\begin{aligned}
(1 - F_x(x)) &= \left(\frac{t_0}{x}\right)^{\alpha_p} \\
\ln(1 - F_x(x)) &= \alpha_p \cdot \ln\left(\frac{t_0}{x}\right) \\
&= \alpha_p \cdot (\ln(t_0) - \ln(x)) \\
&= -\alpha_p \cdot \ln(x) + \alpha_p \cdot \ln(t_0)
\end{aligned}$$

Daraus ergeben sich die Werte der Geradengleichung (nach (JKB70a)) zu:

$$\begin{aligned}
y &= \ln(1 - F_x(x)) \\
x &= \ln(x) \\
a &= -\alpha_p \\
b &= \alpha_p \cdot \ln(t_0)
\end{aligned}$$

Sind der Anstieg und der Schnittpunkt mit der Ordinatenachse mittels linearer Regression bestimmt, berechnen sich die Parameter der Paretoverteilung zu

$$\hat{\alpha}_p = -a \text{ und } \hat{t}_0 = e^{\left(\frac{b}{\hat{\alpha}_p}\right)}. \tag{2.5}$$

Die MATLAB-Funktion *LSE_pareto* stellt eine Implementierung des LSE der Paretoverteilung dar.

2.2.3 Anwendung für die Weibullverteilung

Zum Abschluß wird die Herleitung des LSE für die Weibullverteilung betrachtet. Ausgangspunkt stellt wiederum die Verteilungsfunktion dar (vgl. [1.19]). Die lineare Transformation berechnet sich nach:

$$\begin{aligned}
(1 - F_x(x)) &= e^{-(\frac{x}{\beta})^{\alpha_w}} \\
\ln(1 - F_x(x)) &= -\left(\frac{x}{\beta}\right)^{\alpha_w} \\
-\ln(1 - F_x(x)) &= \left(\frac{x}{\beta}\right)^{\alpha_w} \\
\ln(-\ln(1 - F_x(x))) &= \alpha_w \cdot \ln\left(\frac{x}{\beta}\right) \\
&= \alpha_w \cdot \ln(x) - \alpha_w \cdot \ln(\beta)
\end{aligned}$$

Daraus ergibt sich für die Geradengleichung (vgl. (JKB70a))

$$\begin{aligned}
y &= \ln(-\ln(1 - F_x(x))) \\
x &= \ln(x) \\
a &= \alpha_w \\
b &= -\alpha_w \cdot \ln(\beta).
\end{aligned}$$

Mittels linearer Regression sind die Werte für n und m zu bestimmen. Die Least Squares Estimator der Weibullverteilung sind dann gegeben mit:

$$\hat{\alpha}_w = a \text{ und } \hat{\beta} = e^{-\frac{b}{\alpha_w}} \tag{2.6}$$

In der Funktion *LSE_ weibull* ist eine Realisierung zu finden.

2.3 Methode der Momente

Die Methode der Momente (auch: *Momentenmethode*) ist die historisch älteste Technik zur Schätzung von Verteilungsparametern. Sie wird ausführlich in (RS02) und (Sch01) erläutert. Die Momentenmethode beruht auf dem Prinzip, die theoretischen Momente einer Verteilung durch die entsprechenden Stichprobenmomente zu ersetzen. Da die Momente von den unbekannten Parametern abhängen, ergibt sich hier ein Gleichungssystem, dessen Lösung die Parameterschätzung liefert. Die Anzahl der zu betrachtenden Momente ergibt sich aus der Anzahl der zu schätzenden Parameter.
In der folgenden Tabelle (2.1) sind die Momente der zu betrachtenden Verteilungen zusammenfassend dargestellt. Nach (Sch01) liefert diese Methode meist keine entgültigen

Verteilung	Erwartungswert (EX)	Varianz ($Var\ X$)
Exponential	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Lognormal	$e^{\mu_L + \frac{{\sigma_L}^2}{2}}$	$e^{\left(2\mu_L + {\sigma_L}^2\right)} \cdot \left(e^{{\sigma_L}^2} - 1\right)$
Pareto	$\frac{\alpha_p t_0}{\alpha_p - 1}$	$\frac{\alpha_p {t_0}^2}{(\alpha_p - 2)(\alpha_p - 1)^2}$
Weibull	$\beta \cdot \Gamma\left(1 + \frac{1}{\alpha_w}\right)$	$\beta^2 \left(\Gamma\left(1 + \frac{2}{\alpha_w}\right) - \Gamma^2\left(1 + \frac{1}{\alpha_w}\right)\right)$

Tabelle 2.1: Momente der Verteilungen

Werte, sondern Startwerte für weitere Verfahren, wie dem Maximum Likelihood Estimator. Im allgemeinen ist die Genauigkeit dieser Methode stark abhängig von der Anzahl der Werte in der Messreihe. Je mehr Werte vorhanden sind, desto genauer können die Momente der Grundgesamtheit bestimmt werden.

2.3.1 Anwendung für die Exponentialverteilung

Da die Exponentialverteilung nur einen Parameter besitzt, ist für die Methode der Momente nur der Erwartungswert zu betrachten. Dieser ist gegeben mit

$$EX = \frac{1}{\lambda}.$$

Der Momentenschätzer der Exponentialverteilung ist demnach definiert als

$$\hat{\lambda} = \frac{1}{EX} = \frac{1}{\mu}. \tag{2.7}$$

In den folgenden Abschnitten werden die Momentenschätzer zweiparametriger Verteilungen betrachtet.

2.3.2 Anwendung für die Lognormalverteilung

Im folgenden wird die Methode der Momente auf die Lognormalverteilung angewandt. Ausgehend von den ersten zwei Momenten der Lognormalverteilung

$$EX = e^{\left(\mu_L + \frac{{\sigma_L}^2}{2}\right)} \text{ und } Var\ X = e^{\left(2\mu_L + {\sigma_L}^2\right)} \cdot \left(e^{{\sigma_L}^2} - 1\right), \tag{2.8}$$

ist die Gleichung für den Erwartungswert nach ${\sigma_L}^2$ umzustellen

$${\sigma_L}^2 = 2\ln(EX) - 2\mu_L. \tag{2.9}$$

Einsetzen in die Gleichung für die Varianz und umstellen nach μ_L ergibt:

$$\begin{aligned} Var\ X &= e^{(2\mu_L + 2\ln(EX) - 2\mu_L)} \cdot \left(e^{(2\ln(EX) - 2\mu_L)} - 1\right) \\ &= e^{(2\ln(EX))} \cdot e^{(2\ln(EX) - 2\mu_L)} - e^{(2ln(EX))} \\ e^{(2\ln(EX) - 2\mu_L)} &= \frac{Var\ X}{e^{(2\ln(EX))}} + 1 \end{aligned}$$

Der Momentenschätzer für den Parameter μ_L ist so definiert als

$$\hat{\mu}_L = \frac{2\ln(EX) - \ln\left(\frac{Var\ X}{e^{(2\ln(EX))}} + 1\right)}{2}. \tag{2.10}$$

Mit dieser Gleichung ergibt sich der Momentenschätzer für den Parameter σ_L zu

$$\hat{\sigma}_L^2 = 2\ln(EX) - 2\hat{\mu}_L. \tag{2.11}$$

Eine Implementierung ist in der Funktion *momente_ lognormal* zu finden.

2.3.3 Anwendung für die Paretoverteilung

Ausgehend von den ersten beiden Momenten der Paretoverteilung

$$EX = \frac{\alpha_p t_0}{\alpha_p - 1} \text{ und } Var\ X = \frac{\alpha_p {t_0}^2}{(\alpha_p - 2) \cdot (\alpha_p - 1)^2} \tag{2.12}$$

werden im folgenden die Momentenschätzer beschrieben. Hier ist zunächst die Gleichung für den Erwartungswert zu quadrieren und wie folgt umzustellen:

$${\alpha_p}^2 {t_0}^2 = EX^2 \cdot (\alpha_p - 1)^2. \tag{2.13}$$

Durch einsetzen in die Gleichung für die Varianz ergibt sich

$$\begin{aligned} Var\ X &= \frac{EX^2 \cdot (\alpha_p - 1)^2}{\alpha_p \cdot (\alpha_p - 2) \cdot (\alpha_p - 1)^2} \\ \frac{Var\ X}{EX^2} &= \frac{1}{\alpha_p \cdot (\alpha_p - 2)} \\ 0 &= \alpha_p^2 - 2\alpha_p - \frac{EX^2}{Var\ X} \\ \alpha_{p(1,2)} &= -\frac{-2}{2} \pm \sqrt{\frac{-2^2}{4} + \frac{EX^2}{Var\ X}} \\ \hat{\alpha}_{p(1,2)} &= 1 \pm \sqrt{1 + \frac{EX^2}{Var\ X}}. \end{aligned} \tag{2.14}$$

Die obige Gleichung führt auf zwei Schätzer für den Parameter α. Um den richtigen Schätzer zu ermitteln, sind beide Lösungen gegen t_0 zu prüfen mit:

$$\hat{t}_{0(1,2)} = \frac{EX \cdot \left(\hat{\alpha}_{p(1,2)} - 1\right)}{\hat{\alpha}_{p(1,2)}}. \tag{2.15}$$

Weiterhin muß gelten

$$\hat{t}_0 \approx min(x_i). \tag{2.16}$$

Die MATLAB-Funktion *momente_pareto* ist wiederum als Realisierung zu sehen.

2.3.4 Anwendung für die Weibullverteilung

Zur Schätzung der Parameter der Weibullverteilung mit Hilfe der Methode der Momente wird im folgenden auf die hierfür benötigten Gleichungen eingegangen. Die ersten zwei Momente sind gegeben mit

$$EX = \beta \cdot \Gamma\left(1 + \frac{1}{\alpha_w}\right) \text{ und } Var\ X = \beta^2 \left(\Gamma\left(1 + \frac{2}{\alpha_w}\right) - \Gamma^2\left(1 + \frac{1}{\alpha_w}\right)\right). \tag{2.17}$$

Durch quadrieren der Gleichung für den Erwartungswert und umstellen nach β ergibt sich:

$$\beta^2 = \frac{EX^2}{\Gamma^2\left(1 + \frac{1}{\alpha_w}\right)}$$

Mit einsetzen des obigen Ausdrucks in die Gleichung für die Varianz folgt:

$$\begin{aligned} Var\ X &= \frac{EX^2}{\Gamma^2\left(1+\frac{1}{\alpha_w}\right)} \cdot \left(\Gamma\left(1+\frac{2}{\alpha_w}\right) - \Gamma^2\left(1+\frac{1}{\alpha_w}\right)\right) \\ Var\ X &= \frac{EX^2 \cdot \Gamma\left(1+\frac{2}{\alpha_w}\right)}{\Gamma^2\left(1+\frac{1}{\alpha_w}\right)} - \frac{EX^2 \cdot \Gamma^2\left(1+\frac{1}{\alpha_w}\right)}{\Gamma^2\left(1+\frac{1}{\alpha_w}\right)} \\ \frac{Var\ X - EX^2}{EX^2} &= \frac{\Gamma\left(1+\frac{2}{\hat{\alpha}_w}\right)}{\Gamma^2\left(1+\frac{1}{\hat{\alpha}_w}\right)}. \end{aligned} \tag{2.18}$$

Hierfür gibt es jedoch keine geschlossene Lösung und so ist der Schätzer für den Parameter α_w numerisch zu bestimmen. Der Momentenschätzer für den Parameter β ergibt sich zu:

$$\hat{\beta} = \frac{EX}{\Gamma\left(1+\frac{1}{\hat{\alpha}_w}\right)}. \tag{2.19}$$

Implementiert wurde der Momentenschätzer in der Funktion *momente_weibull*.

2.3.5 Übersicht der Momentenschätzer

In der folgenden Tabelle 2.2 werden noch einmal die Momentenschätzer der Verteilungen zusammenfassend dargestellt.

Verteilung	Momentenschätzer
Exponential	$\hat{\lambda} = \frac{1}{EX} = \frac{1}{\mu}$
Lognormal	$\hat{\mu}_L = \frac{1}{2} \cdot \left(2\ln(\mu) - \ln\left(\frac{\sigma^2}{e^{(2\ln(\mu))}} + 1\right)\right),\ \hat{\sigma}_L^2 = 2\ln(\mu) - 2\hat{\mu}_L$
Pareto	$\hat{\alpha_{p}}_{1,2} = 1 \pm \sqrt{1 + \frac{EX^2}{Var\ X}},\ \hat{t}_{0(1,2)} = \frac{EX \cdot \left(\hat{\alpha}_{p(1,2)} - 1\right)}{\hat{\alpha}_{p(1,2)}}$
Weibull	$\frac{Var\ X - EX^2}{EX^2} = \frac{\Gamma\left(1+\frac{2}{\hat{\alpha}_w}\right)}{\Gamma^2\left(1+\frac{1}{\hat{\alpha}_w}\right)},\ \hat{\beta} = \frac{EX}{\Gamma\left(1+\frac{1}{\hat{\alpha}_w}\right)}$

Tabelle 2.2: Übersicht der Momentenschätzer

2.4 Maximum Likelihood Estimator

Aus statistischer Sicht ist nach (HEK93) und (RS02) der Maximum Likelihood Schätzer eine der robustesten Parameterschätztechniken. Die Idee hinter MLE ist das Errechnen der wahrscheinlichsten Werte für die Parameter einer ausgewählten Verteilung. Die Parameter werden so geschätzt, daß deren Werte die Daten am besten beschreiben. Ausgehend von einer kontinuierlichen, zufälligen Variablen x ist die Wahrscheinlichkeitsdichteverteilungsfunktion gegeben mit

$$f(x; \theta_1, \theta_2, \ldots, \theta_k).$$

Hierbei sind $\theta_1, \theta_2, \ldots, \theta_k$ die k zu bestimmenden Parameter. Des weiteren wird von N unabhängigen Beobachtungen $x_1, x_2, \ldots, x_N$ der Zufallsvariable ausgegangen. Die zugehörige Likelihood-Funktion ist definiert als:

$$\mathcal{L}(x_1, x_2, \ldots, x_N | \theta_1, \theta_2, \ldots, \theta_k) = \mathcal{L} = \prod_{i=1}^{N} f(x_i; \theta_1, \theta_2, \ldots, \theta_k). \tag{2.20}$$

Um die Berechnung der Maximum Likelihood Schätzer zu vereinfachen, wird die log-likelihood Funktion berechnet, die gegeben ist mit:

$$\Lambda = \ln \mathcal{L} = \sum_{i=1}^{N} \ln(f(x_i; \theta_1, \theta_2, \ldots, \theta_k)). \tag{2.21}$$

Wird die Funktion $\mathcal{L}$ bzw. Λ in Abhängigkeit von θ maximiert, so ergibt das die Maximum-Likelihood Schätzung für den Parameter θ. Die Maximierung erfolgt mittels partieller Differenzierung nach

$$\frac{\partial \Lambda}{\partial \theta_j} = 0 \text{ mit } j = 0, 1, \ldots, k$$

und durch Berechnung der Nullstellen.
Am Beispiel der Normalverteilung werden im folgenden die Maximum Likelihood Schätzer für die Parameter μ und σ ermittelt. Ausgehend von der Dichtefunktion

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.22}$$

ergibt sich die zugehörige Likelihood Funktion zu

$$\begin{aligned} \mathcal{L} &= \mathcal{L}(x_1, \ldots, x_N | \mu, \sigma) \\ &= \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \qquad (2.23) \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^N} \cdot e^{-\frac{1}{2}\sum\limits_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2}. \qquad (2.24) \end{aligned}$$

Durch logarithmische Transformation berechnet sich die loglikelihood Funktion nach:

$$\begin{aligned}\Lambda &= \ln(\mathcal{L}) = -N \cdot \ln\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2 \\ &= -N \cdot \ln(\sigma) - \frac{N}{2}\cdot\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{x_i-\mu}{\sigma}\right)^2\end{aligned}$$

Für den Maximum Likelihood Schätzer $\hat{\sigma}$ ist die log-likelihood Funktion partiell nach σ zu differenzieren

$$\begin{aligned}\frac{\partial\Lambda}{\partial\sigma} &= -\frac{N}{\sigma} - \frac{1}{2}\sum_{i=1}^{N} -2\frac{(x_i-\mu)^2}{\sigma^3} \\ &= -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2\end{aligned}$$

und die Nullstelle zu berechnen.

$$\begin{aligned}0 &= -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 \\ N\cdot\sigma^2 &= \sum_{i=1}^{N}(x_i-\mu)^2 \\ \hat{\sigma} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}\end{aligned} \tag{2.25}$$

Es ist zu erkennen, daß $\hat{\sigma}^2$ dem zweiten Moment der Normalverteilung entspricht. Der Maximum Likelihood Schätzer $\hat{\mu}$ ergibt sich zu:

$$\begin{aligned}\frac{\partial\Lambda}{\partial\mu} &= -\frac{1}{2}\sum_{i=1}^{N}\frac{-2x_i+2\mu}{\sigma^2} \\ &= \sum_{i=1}^{N}\frac{x_i-\mu}{\sigma^2} \\ 0 &= \frac{-N\cdot\mu}{\sigma^2}\sum_{i=1}^{N}\frac{x_i}{\sigma^2} \\ \frac{N\cdot\mu}{\sigma^2} &= \sum_{i=1}^{N}\frac{x_i}{\sigma^2} \\ \hat{\mu} &= \frac{1}{N}\sum_{i=1}^{N}x_i\end{aligned} \tag{2.26}$$

Der MLE für den Parameter μ ist das erste Moment der Normalverteilung.
Im allgemeinen ist der MLE zur Bestimmung von Verteilungsparametern anderen Methoden vorzuziehen, da dieser (nach (RS02)) die genauesten Werte liefert. Inwiefern diese

Aussage zutreffend ist, wird in Kapitel 5 ausführlich dargestellt. Ein weiterer Vorteil des MLE liegt in der Möglichkeiten zur Angabe von Konfidenzintervallen, auf die in dieser Arbeit nicht näher eingegangen wird.
Für die hier betrachteten Verteilungen werden die Herleitungen der Maximum Likelihood Estimator in den folgenden Abschnitten gezeigt, beginnend mit der Exponentialverteilung.

2.4.1 Anwendung für die Exponentialverteilung

In diesem Abschnitt wird der Maximum Likelihood Estimator für den Paramter λ der Exponentialverteilung näher betrachtet. Ausgehend von der Dichtefunktion (Gleichung [1.7]) berechnet sich die log-likelihood Funktion zu:

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}(x_1, \ldots, x_N|\lambda) \\
&= \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \lambda \cdot e^{-\lambda x_i} \\
\Lambda &= \ln(\mathcal{L}) \\
&= \sum_{i=1}^{N} \ln \lambda - \lambda x_i \\
&= N \cdot \ln \lambda - \lambda \sum_{i=1}^{N} x_i. \qquad (2.27)
\end{aligned}$$

(2.28)

Durch partielles Differenzieren und Berechnen der Nullstellen berechnet sich der Maximum Likelihood Estimator für den Parameter λ nach (RSC04) gegeben mit:

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \lambda} &= \frac{N}{\lambda} - \sum_{i=1}^{N} x_i \\
0 &= \frac{N}{\lambda} - \sum_{i=1}^{N} x_i \\
\frac{1}{\lambda} &= \frac{1}{N} \sum_{i=1}^{N} x_i \\
\hat{\lambda} &= \frac{N}{\sum_{i=1}^{N} x_i} = \frac{1}{\mu}. \qquad (2.29)
\end{aligned}$$

Es ist zu erkennen, daß der MLE für den Parameter λ dem Erwartungswert der Exponentialverteilung entspricht. Die Methode der Momente und der MLE gehen hier von dem gleichem Zusammenhang aus.

2.4.2 Anwendung für die Lognormalverteilung

In diesem Abschnitt wird die Herleitung der Maximum Likelihood Estimator der Lognormalverteilung näher betrachtet. Die Likelihood Funktion und die loglikelihood Funktion berechnen sich nach:

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}(x_1, \ldots, x_N | \mu_L, \sigma_L) \\
&= \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \frac{1}{x_i \sigma_L \sqrt{2\pi}} e^{-\frac{(\ln(x_i)-\mu_L)^2}{2{\sigma_L}^2}} \qquad (2.30) \\
&= e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\ln(x_i)-\mu_L}{\sigma_L}\right)^2} \cdot \prod_{i=1}^{N} \frac{1}{\left(x_i \sigma_L \sqrt{2\pi}\right)} \qquad (2.31) \\
\Lambda &= \ln(\mathcal{L}) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\ln(x_i)-\mu_L}{\sigma_L}\right)^2 + \sum_{i=1}^{N} \ln\left(\frac{1}{x_i \sigma_L \sqrt{2\pi}}\right) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\ln(x_i)-\mu_L}{\sigma_L}\right)^2 + \sum_{i=1}^{N} -ln(x_i) - \ln(\sigma_L) - \frac{1}{2}ln(2\pi) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\ln(x_i)-\mu_L}{\sigma_L}\right)^2 - N \cdot \ln(\sigma_L) - \frac{N}{2}\ln(2\pi) - \sum_{i=1}^{N}\ln(x_i).
\end{aligned}$$

Durch partielles Differenzieren und Berechnung der Nullstellen ergibt sich der MLE für σ_L zu

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \sigma_L} &= -\frac{N}{\sigma_L} - \frac{1}{2}\sum_{i=1}^{N} -2 \cdot \frac{(\ln(x_i)-\mu_L)^2}{{\sigma_L}^3} \\
0 &= -\frac{N}{\sigma_L} + \sum_{i=1}^{N} \frac{(\ln(x_i)-\mu_L)^2}{{\sigma_L}^3} \\
N \cdot {\sigma_L}^2 &= \sum_{i=1}^{N} (\ln(x_i)-\mu_L)^2 \\
\hat{\sigma}_L &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\ln(x_i)-\mu_L)^2}. \qquad (2.32)
\end{aligned}$$

Der Maximum Likelihood Schätzer für μ_L ist somit gegeben mit:

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \mu_L} &= -\frac{1}{2}\sum_{i=1}^{N}\frac{-2\ln(x_i)+2\mu_L}{{\sigma_L}^2} \\
&= \sum_{i=1}^{N}\left(\frac{\ln(x_i)}{{\sigma_L}^2}-\frac{\mu_L}{{\sigma_L}^2}\right) \\
0 &= -\frac{N\cdot\mu_L}{{\sigma_L}^2}+\sum_{i=1}^{N}\frac{\ln(x_i)}{{\sigma_L}^2} \\
N\cdot\mu_L &= \sum_{i=1}^{N}\ln(x_i) \\
\hat{\mu}_L &= \frac{1}{N}\sum_{i=1}^{N}\ln(x_i) \qquad (2.33)
\end{aligned}$$

Die hier gezeigten Schätzer sind denen der Normalverteilung (vgl. Abschnitt 2.4) sehr ähnlich. Bei der praktischen Anwendung werden das arithmetische Mittel und die Standardabweichung der logarithmierten Daten ermittelt. Auf diese Weise berechnet die Funktion *lognfit* in MATLAB die Maximum Likelihood Estimator der Lognormalverteilung.

2.4.3 Anwendung für die Paretoverteilung

Ausgehend von der Dichtefunktion der Paretoverteilung (vgl. Gleichung [1.14]) werden in diesem Abschnitt die Maximum Likelihood Estimator der Parameter der Paretoverteilung näher betrachtet. Die log-likelihood Funktion ist gegeben mit

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}\left(x_1,\ldots,x_N|\alpha_p,t_0\right) \\
&= \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N}\frac{\alpha_p}{x_i}\cdot\left(\frac{t_0}{x_i}\right)^{\alpha_p} \qquad (2.34) \\
\Lambda &= \ln\mathcal{L} \\
&= \sum_{i=1}^{N}\ln\left(\frac{\alpha_p}{x_i}\right)+\alpha_p\cdot\ln\left(\frac{t_0}{x_i}\right) \\
&= \sum_{i=1}^{N}\left(\ \ln(\alpha_p)-\ln(x_i)+\alpha_p\cdot\ln(t_0)-\alpha_p\cdot\ln(x_i)\ \right) \\
&= N\cdot\ln(\alpha_p)-\sum_{i=1}^{N}\ln(x_i)+\alpha_p\cdot N\cdot\ln(t_0)-\alpha_p\cdot\sum_{i=1}^{N}\ln(x_i).
\end{aligned}$$

Der MLE für den Parameter α_p berechnet sich durch partielles Differenzieren und Berechnung der Nullstellen, wie folgt (vgl. (JKB70a)):

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \alpha_p} &= \frac{N}{\alpha_p} + N \cdot \ln(t_0) - \sum_{i=1}^{N} \ln(x_i) = 0 \\
\frac{N}{\alpha_p} &= \sum_{i=1}^{N} \ln(x_i) - N \cdot \ln(t_0) = \sum_{i=1}^{N} \ln(x_i) - N \cdot \frac{1}{N} \sum_{i=1}^{N} \ln(t_0) \\
\hat{\alpha}_p &= \frac{N}{\sum\limits_{i=1}^{N} \ln\left(\frac{x_i}{t_0}\right)}.
\end{aligned} \tag{2.35}$$

Der Maximum Likelihood Schätzer für den Parameter t_0 der Verteilung ist nach (JKB70a) definiert durch:

$$\hat{t}_0 = \min x_i. \tag{2.36}$$

Dies stimmt mit der Dichtefunktion der Paretoverteilung überein, die nicht definiert ist für $x < t_0$. Hier offenbart sich jedoch ein Nachteil der Schätzung mittels MLE. Sowohl α_p als auch t_0 bestimmen die Form der Verteilung. Ist t_0 vorgegeben, kann die Form nur noch mit dem Parameter α_p angepasst werden. Eine Implementierung stellt die Funktion *MLE_pareto* dar.

2.4.4 Anwendung für die Weibullverteilung

Im folgenden werden die Maximum Likelihood Estimator für die Parameter der Weibullverteilung vorgestellt. Zunächst folgt die Betrachtung des MLE für den Parameter β.

$$\begin{aligned}
\mathcal{L} &= \mathcal{L}\left(x_1, \ldots, x_N | \alpha_w, \beta\right) \\
&= \prod_{i=1}^{N} f(x_i) = \prod_{i=1}^{N} \frac{\alpha_w}{\beta} \left(\frac{x_i}{\beta}\right)^{\alpha_w - 1} \cdot exp\left(-\left(\frac{x_i}{\beta}\right)^{\alpha_w}\right)
\end{aligned} \tag{2.37}$$

$$\begin{aligned}
\Lambda &= \ln \mathcal{L} \\
&= \sum_{i=1}^{N} \left(\ln\left(\frac{\alpha_w}{\beta}\right) + (\alpha_w - 1) \cdot \ln\left(\frac{x}{\beta}\right) - \left(\frac{x}{\beta}\right)^{\alpha_w}\right) \\
&= N \cdot \ln\left(\frac{\alpha_w}{\beta}\right) + \alpha_w \cdot \sum_{i=1}^{N} \ln\left(\frac{x_i}{\beta}\right) - \sum_{i=1}^{N} \ln(x_i) + N \cdot \ln(\beta) - \beta^{-\alpha_w} \cdot \sum_{i=1}^{N} x^{\alpha_w}
\end{aligned} \tag{2.38}$$

$$\begin{aligned}
\frac{\partial \Lambda}{\partial \beta} &= -\frac{\alpha_w \cdot N}{\beta} + \alpha_w \cdot \beta^{-\alpha_w - 1} \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w} = 0 \\
\frac{N}{\beta} &= \beta^{-\alpha_w - 1} \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w} \\
\frac{1}{\beta^{-\alpha_w}} &= \frac{1}{N} \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w} \\
\hat{\beta} &= \sqrt[\hat{\alpha}_w]{\frac{1}{N} \cdot \sum_{i=1}^{N} {x_i}^{\hat{\alpha}_w}}
\end{aligned} \tag{2.39}$$

Ausgehend vom Maximum Likelihood Estimator für den Parameter β berechnet sich der Maximum Likelihood Schätzer für den Parameter α_w wie folgt:

$$\begin{aligned}
\Lambda &= \ln \mathcal{L} \\
&= N \cdot \ln\left(\frac{\alpha_w}{\beta}\right) + \alpha_w \cdot \sum_{i=1}^{N} \ln\left(\frac{x_i}{\beta}\right) - \sum_{i=1}^{N} \ln(x_i) + N \cdot \ln(\beta) - \beta^{-\alpha_w} \cdot \sum_{i=1}^{N} x^{\alpha_w} \\
\frac{\partial \Lambda}{\partial \alpha_w} &= \frac{N}{\alpha_w} + \sum_{i=1}^{N} \ln(x_i) - N \cdot \ln(\beta) - \frac{\sum_{i=1}^{N} {x_i}^{\alpha_w} \cdot \ln(x_i)}{\beta^{\alpha_w}} + \frac{\ln(\beta) \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w}}{\beta^{\alpha_w}}
\end{aligned}$$

Durch einsetzen von Gleichung [2.39] ergibt sich:

$$\begin{aligned}
0 &= \frac{N}{\alpha_w} + \sum_{i=1}^{N} ln(x_i) - \frac{N}{\alpha_w} \cdot \ln\left(\frac{1}{N} \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w}\right) - \frac{\sum_{i=1}^{N} {x_i}^{\alpha_w} \cdot \ln(x_i)}{\frac{1}{N} \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w}} + \\
&\quad + \frac{N}{\alpha_w} \cdot \frac{\ln\left(\frac{1}{N} \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w}\right) \cdot \sum_{i=1}^{N} {x_i}^{\alpha_w}}{\cdot \sum_{i=1}^{N} {x_i}^{\alpha_w}} \\
&= \frac{N}{\alpha_w} + \sum_{i=1}^{N} \ln(x_i) - N \cdot \frac{\sum_{i=1}^{N} {x_i}^{\alpha_w} \cdot \ln(x_i)}{\sum_{i=1}^{N} {x_i}^{\alpha_w}} \\
\frac{N}{\alpha_w} &= N \cdot \frac{\sum_{i=1}^{N} {x_i}^{\alpha_w} \cdot \ln(x_i)}{\sum_{i=1}^{N} {x_i}^{\alpha_w}} - \sum_{i=1}^{N} \ln(x_i) \\
\hat{\alpha}_w &= \frac{1}{\frac{\sum_{i=1}^{N} {x_i}^{\hat{\alpha}_w} \cdot \ln(x_i)}{\sum_{i=1}^{N} {x_i}^{\hat{\alpha}_w}} - \frac{1}{N} \cdot \sum_{i=1}^{N} \ln(x_i)}.
\end{aligned} \tag{2.40}$$

Die Lösungen für die Maximum Likelihood Schätzer stimmen mit den Angaben in (JKB70a) überein. Die in MATLAB zur Verfügung stehende Funktion *wblfit* stellt eine Implementierung der hier vorgestellten Schätzer der Weibullparameter dar.

2.4.5 Übersicht der Maximum Likelihood Schätzer

In Tabelle 2.3 werden die MLE der Verteilungen zusammengefasst. Eine ähnliche Tabelle

Verteilung	Maximum Likelihood Schätzer
Exponential	$\hat{\lambda} = \frac{1}{\overline{EX}} = \frac{1}{\mu}$
Lognormal	$\hat{\sigma}_L = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\ln(x_i) - \mu_L)^2}, \;\; \hat{\mu}_L = \frac{1}{N}\sum_{i=1}^{N}\ln(x_i)$
Pareto	$\hat{\alpha}_p = N \cdot \left(\sum_{i=1}^{N} \ln\left(\frac{x_i}{t_0}\right)\right)^{-1}, \;\; \hat{t}_0 = \min x_i$
Weibull	$\hat{\beta} = \sqrt[\hat{\alpha}_w]{\frac{1}{N}\cdot\sum_{i=1}^{N} {x_i}^{\hat{\alpha}_w}}, \;\; \hat{\alpha}_w = \left[\left(\sum_{i=1}^{N} {x_i}^{\hat{\alpha}_w}\cdot\ln(x_i)\right)\cdot\left(\sum_{i=1}^{N} {x_i}^{\hat{\alpha}_w}\right)^{-1} - \frac{1}{N}\cdot\sum_{i=1}^{N}\ln(x_i)\right]^{-1}$

Tabelle 2.3: Übersicht der Maximum Likelihood Schätzer

ist in (Fel01) zu finden. Die hier gezeigten Lösungen entsprechen den Ergebnissen in der genannten Veröffentlichung.

Kapitel 3

Anpassungskriterien

Nach (Pax94) ist auf dem Gebiet der Statistik keine exakte Beschreibung für sehr große Datenmengen möglich. Bei der Untersuchung von Meßreihen handelt es sich immer nur um die Betrachtung einer Stichprobe der Grundgesamtheit. So ist davon auszugehen, daß auch die geschätzten Parameter nicht statistich exakt ermittelt werden können. Es gilt herauszufinden, wie genau das Modell mit der untersuchten Verteilung übereinstimmt. Betrachtet wird somit die Güte der Anpassung (engl.: *Goodness-of-Fit*).
In den folgenden Abschnitten werden einige der gebräuchlichsten Verfahren näher betrachtet. Dabei wird auf deren Vor- und Nachteile eingegangen. Weiterhin wird geprüft, ob sie in Zusammenhang mit dieser Arbeit anwendbar sind.
Nach (AS86) sind auf dem Gebiet der Statistik formal folgende Hypothesen zu überprüfen:

$$H_0 : F_X(x) = F_0(x|\hat{\theta}) \text{ gegen } H_1 : F_X(x) \neq F_0(x|\hat{\theta})$$

Hier ist F_X die empirische und F_0 die theoretische Verteilungsfunktion. Die theoretische Verteilungsfunktion wird beschrieben durch die geschätzten Parameter $\hat{\theta}$.

3.1 Kolmogorov-Smirnov Anpassungstest

Mit dieser Methode wird getest, ob die Verteilung einer Zufallsvariablen einer bestimmten theoretischen Verteilung zuzuordnen ist. Es gilt die Annahme, daß die Werte der Stichprobe einer stetigen Verteilung folgen. Eine Anwendung auf die Poissonverteilung ist demnach nicht möglich. Es ist naheliegend, die absoluten Differenzen zwischen der empirischen Verteilung $F_X(x)$ und der theoretischen Verteilung $F_0(x)$

$$d = |F_X(x) - F_0(x)| \tag{3.1}$$

zu ermitteln. Dies wird in Abbildung (3.1) verdeutlicht. So ist der Kolomogorov-Smirnov

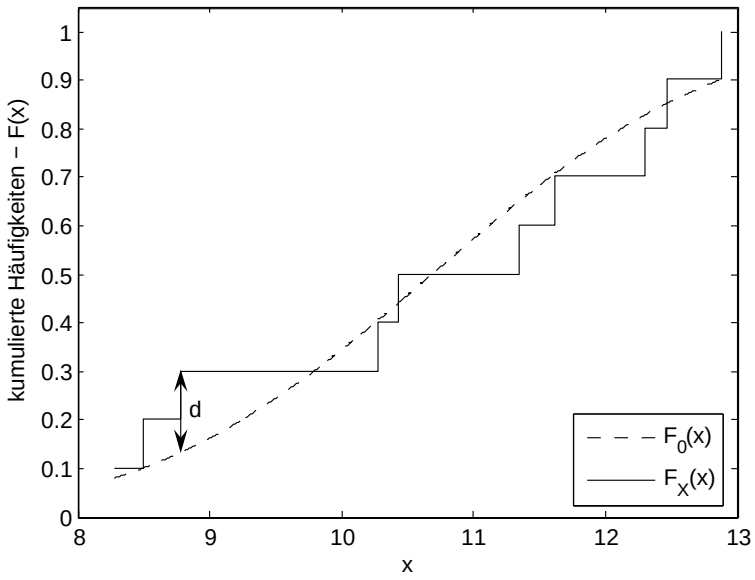


Abbildung 3.1: Kolmogorov-Smirnov Anpassungstest am Beispiel einer Normalverteilung

Anpassungstest (kurz: *K-S-Test*) nach (Dut03) und (AS86) definiert als:

$$d_{max} = \max_{1 \le i \le N} |F_X(x_i) - F_0(x_i)| \tag{3.2}$$

Der K-S-Test beinhaltet anschließend den Vergleich des Wertes d_{max} mit einem entsprechenden Wert in einer Tabelle, abhängig vom Signifikanzniveau α und der Anzahl der Werte der empirischen Verteilung N. Gilt

$$d_{N,\alpha} > d_{max} \tag{3.3}$$

kann die Nullhypothese H_0 angenommen werden. Eine entsprechende Tabelle ist in (Dut03) zu finden. Vorteil dieser Methode ist, daß hier keine Klassen zu bilden sind, was zu einem Informationsverlust führen würde. Der Nachteil ergibt sich aus der Tabelle, die nur für $N \le 40$ Vergleichswerte vorsieht. In den hier untersuchten Meßreihen wurden jedoch einige tausend Werte betrachtet. Es ist anzunehmen, daß die Anwendung der Näherungsformeln für $N > 40$ bei deutlich größeren Werten für N zu Fehlern führt.

Da das Auftreten dieses Fehlers nicht ausgeschlossen werden kann, werden in folgenden Betrachtungen nur die maximalen Differenzen d_{max} betrachtet, ohne den Vergleich mit dem Wert in der Tabelle anzustellen. Paxson selbst hat in (Pax94) die Anwendung des Kolmogorov-Smirnov Anpassungstests in einer seiner früheren Arbeiten als ungeeignet bezeichnet. Auch hier wurde Netzwerkverkehr untersucht.

3.2 χ^2-Anpassungstest

Der χ^2-Anpassungstest wird in (Pax94) und (HEK93) ausführlich vorgestellt. Nach (Sch01) eignet sich der wohl bekannteste Anpassungstest auch für große Datenmengen. Überprüft werden hier die in Abschnitt 3 aufgestellten Hypothesen. Hierzu werden die empirischen und theoretischen absoluten Häufigkeiten miteinander verglichen. Ausgehend von n Beobachtungen einer Zufallsvariable Y, sei Z die zu testende theoretische Verteilung. Es folgt die Unterteilung der theoretischen Verteilung in N Klassen. Es muß gelten:

$$N \geq 6 \text{ und } n \geq 30$$

Die zugehörige Teststatistik ist definiert mit

$$\chi^2 = \sum_{i=1}^{N} \frac{(Y_i - n \cdot p_i)^2}{n \cdot p_i}. \tag{3.4}$$

Hierbei entspricht p_i dem Anteil der Verteilung Z, der der i-ten Klasse zugeordnet werden kann. Y_i gibt die Anzahl der Beobachtungen in Y an, die in die i-te Klasse fallen. Abhängig vom Signifikanzniveau α und von der Anzahl der Freiheitsgrade *df* ist das Ergebnis mit dem Wert in einer entsprechenden Tabelle zu vergleichen. Die Nullhypothese ist zu verwerfen, falls gilt:

$$\chi^2 > \chi^2{}_{df;1-\alpha} \tag{3.5}$$

mit

$$df = N - 1 - Est. \tag{3.6}$$

Est gibt die Anzahl der geschätzten Parameter an. In dieser Arbeit findet das Verfahren keine Verwendung, da aggregierte Ankunftsprozesse betrachtet werden. Die Aggregation von Zwischenankunftszeiten entspricht bereits einer Art der Klassenbildung. Die Tabelle sieht des weiteren keine Werte für einige tausend Klassen vor. In (Pax94) wird der λ^2-Anpassungstest vorgestellt, der keinen Vergleich mit den Werten in einer Tabelle vorsieht.

3.3 λ^2-Diskrepanzwert

Hierbei handelt es sich um einen modifizierten χ^2-Anpassungstest. Um unabhängig von den Werten einer Tabelle zu sein, wird in (Pax94) der χ^2-Diskrepanzwert eingeführt, der gegeben ist durch χ^2/n. Paxson zeigt jedoch das Problem auf, daß χ^2-Diskrepanzwerte nicht verglichen werden können, wenn die Anzahl der Klassen nicht übereinstimmt. Als Lösung dieses Problems schlägt Paxson den λ^2-Diskrepanzwert vor.
Aufbauend auf Abschnitt 3.2 sei $E_i = n \cdot p_i$ die erwartete Häufigkeit für die i-te Klasse. Die Diskrepanz in der i-ten Klasse ist definiert durch $D_i = Y_i - E_i$. Weiterhin ist

$$K = \sum_{i=1}^{N} \frac{D_i}{E_i} \tag{3.7}$$

und

$$\hat{\lambda^2} = \frac{\chi^2 - K - df}{n - 1}. \tag{3.8}$$

Zu vergleichen sind die λ^2-Diskrepanzwerte. Bei der Verwendung des λ^2-Tests sind jedoch einige Aspekte zu beachten. So ist auszuwählen, wie viele Klassen zu betrachten sind. Wird N zu klein gewählt, werden die Diskrepanzen nur grob erkannt. Ist N zu groß, werden auch kleinste Abweichungen betrachtet, die nicht von Interesse sind. In (Pax94) wird als grobe Regel für die Blockgröße

$$\omega = 3,49 \cdot \hat{\sigma}_x n^{-\frac{1}{3}} \tag{3.9}$$

angegeben. Hier sind n die Anzahl der Werte einer Zufallsvariablen, ω die Blockgröße und $\hat{\sigma}_x$ als Näherungswert für die Standardabweichung der Verteilung.
Angewandt wurde dieser Test in (Fel01) und (Pax94) für nicht-aggregierten Netzwerkverkehr. In den Veröffentlichungen ergab sich für die Anzahl der Klassen $5 \leq N \leq 240$. Testberechnungen für den in dieser Arbeit betrachteten aggregierten Netzwerkverkehr ergaben jedoch Klassen der Größe $\omega < 1$.
Auch hier verdeutlicht sich der Nachteil des im vorherigen Abschnitt betrachteten χ^2-Anpassungstest. Aufgrund des Informationsverlustes ist von der Klassenbildung bereits klassierter Daten abzusehen. Im Abschnitt 3.6 werden Kriterien aufgezeigt, die weder eine Klassenbildung, noch Vergleichswerte in Tabellen voraussetzen.

3.4 Quantile-Quantile-Plot

Der Quantile-Quantile-Plot (kurz: *Q-Q-Plot*) ist eine graphische Methode für den Vergleich zweier Verteilungen. Ausgehend von den Merkmalsausprägungen der Verteilungen x_i $[i = 1, \ldots, n]$ und y_j $[j = 1, \ldots, m]$, sind die Quantile $q_x(\alpha)$ gegen die Quantile $q_y(\alpha)$ abzutragen (vgl. Abb. 3.2). So wird zum Beispiel der Median x_{Med} gegen den Median y_{Med} dargestellt. Besitzen beide Verteilungen den gleichen Umfang, d.h. $m = n$, so ist der Q-Q-Plot eine einfache Grafik der geordneten x-Werte gegen die geordneten y-Werte. Ist dieser Fall nicht gegeben, ist über die Quantilränge $\alpha(i)$ nach (Gess93) vorzugehen.

Für die Interpretation gilt: Je näher die Datenpunkte an der ersten Winkelhalbierenden

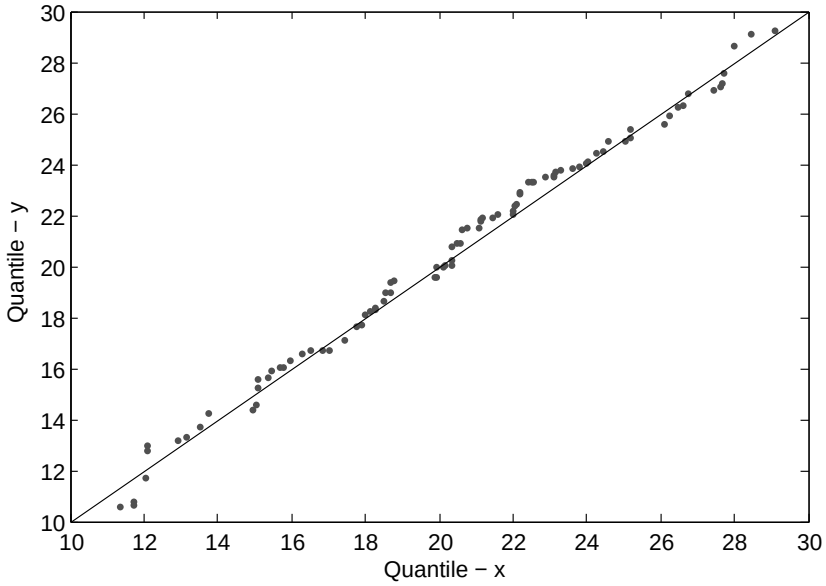


Abbildung 3.2: Q-Q-Plot zweier $\mathcal{N}(20, 5)$-verteilter Zufallsgrößen

(Anstieg $a = 1$) liegen, desto ähnlicher sind sich die Verteilungen. In (LWDW97) wird der Q-Q-Plot verwendet, um Daten einer Meßreihe mit denen einer Simulation zu vergleichen. Eine Implementierung des Q-Q-Plots ist in MATLAB mit der Funktion *qqplot* zu finden. Für den Vergleich der Verteilungsfunktion ist jedoch der P-P-Plot anzuwenden, wie im nächsten Abschnitt beschrieben.

3.5 Probability-Probability-Plot

Ähnlich dem Q-Q-Plot können mit Hilfe des Probabilitiy-Propability Plots (kurz: *P-P-Plot*) zwei Verteilungsfunktionen verglichen werden, wie in (Gess93) beschrieben. Gegeneinander abgetragen werden die kumulierten relativen Häufigkeiten der beiden zu vergleichenden Verteilungen (vgl. Abb. 3.3). Wiederum ausgehend von den Merkmalsausprägun-

gen x_i $[i = 1, \ldots, n]$ und y_j $[j = 1, \ldots, m]$ der Verteilungen, werden die Punkte $F_X(x_i)$ gegen die Werte $F_Y(x_i)$ bzw. die Werte $F_Y(y_i)$ gegen $F_X(y_i)$ dargestellt. Die Interpretati-

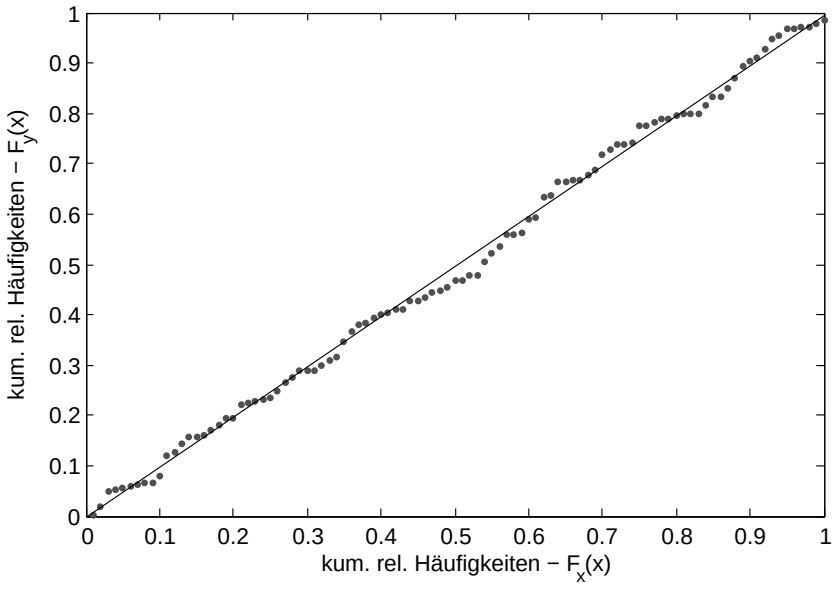


Abbildung 3.3: P-P-Plot am Beispiel von zwei Normalverteilungen

on des P-P-Plots entspricht der des Quantile-Quantile Plots: Je genauer die Datenpunkte der ersten Winkelhalbierenden (Anstieg $a = 1$) entsprechen, desto größer ist der Grad der Anpassung der Verteilungen.
Um die theoretischen mit den empirischen Verteilungen in Kapitel 5 vergleichen zu können, wird diese Darstellungsform angewandt. Die MATLAB-Funktion *ppplot* stellt eine Implementierung dar.

3.6 Weitere Kriterien

Bis auf den P-P-Plot und dem maximalen Abstand zweier Verteilungen d_{max}, werden die in den letzten Abschnitten betrachtet Kriterien aus genannten Gründen nicht verwendet. Die Aussage, daß eine Nullhypothese angenommmen bzw. verworfen werden kann, der klassischen Anpassungstests, ist ungeeignet um eine quantitative Aussage über den Grad der Anpassung zu gestatten.
In den folgenden Abschnitten werden zwei alternative Kriterien vorgestellt, mit denen die Abweichungen zwischen empirischer und theoretischer Verteilung bestimmt werden können. Hierbei handelt es sich um recht einfache Methode, die jedoch ihren Zweck erfüllen, wie in Kapitel 5 zu sehen.

3.6.1 Root Mean Square Error (RMS)

Der Nachteil des maximalen Abstands d_{max} aus Abschnitt 3.1 ist dessen Anfälligkeit auf Abweichungen um den Medianwert der Verteilungen. Es gilt ein Maß für den mittleren Fehler zweier Verläufe zu finden. In diesem Zusammenhang wird der mittlere quadratische Fehler (engl.: *Mean Square Error*, kurz: *MSE*) betrachtet, der die Summe der quadratischen Einzelfehler e_i angibt. Dieser ist nach (BSMM99) definiert mit:

$$MSE \;=\; \frac{1}{N}\sum_{i=1}^{N} e_i^2. \tag{3.10}$$

Werden die Abstände bzw. Differenzen d (vgl. Gleichung [3.1]) für die Einzelfehler eingesetzt ergibt sich der mittlere quadratische Fehler zu:

$$MSE \;=\; \frac{1}{N}\sum_{i=1}^{N} d_i^2 = \frac{1}{N}\sum_{i=1}^{N} \left|F_X(x_i) - F_0(x_i)\right|^2. \tag{3.11}$$

Bei der Berechnung des MSE zwischen zwei Verteilungsfunktionen ist davon auszugehen, daß dessen Werte zwischen 0 und 1 liegen werden. Für einen besseren Vergleich ist die Wurzel zu ziehen. Dies führt zum RMS (*Root Mean Square Error*) nach :

$$RMS \;=\; \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left|F_X(x_i) - F_0(x_i)\right|^2} \tag{3.12}$$

In Kapitel 5 wird der RMS als Maß für die mittlere Abweichung angewandt. Bei steigendem Grad der Anpassung muß dann $RMS \to 0$ gelten.

3.6.2 Korrelationskoeffizient

Der Korrelationskoeffizient wird ausführlich in (Gess93) beschrieben. Angewandt wird der Korrelationskoeffizient häufig, um den Zusammenhang zweier Merkmale quantitativ auszudrücken. Ähnlich dem Q-Q-Plot werden somit die Merkmalsausprägungen
$x_i\ [i = 1, \ldots, n]$ und $y_j\ [j = 1, \ldots, m]$ betrachtet. Für $m = n$ ergibt sich der zugehörige Korrelationskoeffizient nach (BSMM99) zu

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Bei genauerer Betrachtung fällt hier die Ähnlichkeit zu Gleichung [2.2] auf. Wird der Korrelationskoeffizient wie oben beschrieben angewandt, ist er als eine Näherung für den Anstieg der Regressionsgeraden des Q-Q-Plots anzusehen.

Da in Zusammenhang mit dieser Arbeit jedoch nur der P-P-Plot verwendet wird, ist auch der Korrelationskoeffizient dementsprechend anzugleichen. Quantitativ ausgedrückt wird somit der Zusammenhang zwischen den theoretischen und empirischen kumulierten Häufigkeiten ($F_0(x)$ und $F_X(x)$). Dieser Zusammenhang resultiert dann in folgender Gleichung:

$$r = \frac{\sum_{i=1}^{n} \left(F_0(x_i) - \bar{F}_0\right) \left(F_X(x_i) - \bar{F}_X\right)}{\sqrt{\sum_{i=1}^{n} \left(F_0(x_i) - \bar{F}_0\right)^2 \sum_{i=1}^{n} \left(F_X(x_i) - \bar{F}_X\right)^2}}. \tag{3.13}$$

Der Korrelationskoeffizient nimmt Werte aus dem Intervall [-1,1] an. Je stärker der Zusammenhang zwischen zwei Verteilungen ist, desto größer ist der Korrelationskoeffizient. Es ist zu erwarten, daß während der Analyse $r \approx 1$ gilt. Grund hierfür ist das gestellte Ziel der möglichst genauen Anpassung der Verteilungsfunktionen.

Kapitel 4

Methoden zur Schätzung des Hurst-Parameters

In diesem Kapitel werden die in dieser Arbeit verwendeten Schätzverfahren für den Hurst-Parameter vorgestellt. Dieser gilt als Maß für den Grad der Selbstähnlichkeit innerhalb einer betrachteten Meßreihe (vgl. Abschnitt 1.2.3)
Unter anderem in (KFR02) und (LTWW94) wird gezeigt, daß der Hurst-Parameter, je nach verwendetem Schätzverfahren, stark schwanken kann. Ein Grund ist der Grad der Aggregation. In den oben aufgeführten Veröffentlichungen wird vorgeschlagen, mehrere Verfahren zu verwenden, um konkrete Aussagen über den Grad der Selbstähnlichkeit zu ermöglichen.
Das Problem des Hurst-Parameter, daß dessen Wert nur geschätzt werden kann, bleibt bestehen. Bei den hier vorgestellten Methoden handelt es sich um graphische Schätzverfahren. Obwohl es sich dabei um die gebräuchlichsten Methoden handelt, besteht das Problem, daß ein subjektiver Einfluß nicht ausgeschlossen werden kann, wenn der Regressionsbereich zu definieren ist.

4.1 Rescaled Adjusted Range-Statistik

Mit Hilfe der Rescaled Adjusted Range-Statistik (kurz: *R/S-Statistik*) beschrieb Hurst 1955 eine Gesetzmäßigkeit, die in Meßreihen mit Langzeitabhängigkeit auftritt. Die Gesetzmäßigkeit ist als Hurst-Effekt bzw. Hurst's Law bekannt. Untersucht wird hierbei das Verhalten der Meßreihe bei verschiedenen Aggregationsstufen.
Ausgehend von einer Meßreihe $(X_k : k = 1, 2, \ldots, n)$ ist die R/S-Statistik nach (LTWW93) und (Bor01) wie folgt definiert:

$$\frac{R(n)}{S(n)} = \frac{max(0, W_1, W_2, \ldots, W_n) - min(0, W_1, W_2, \ldots, W_n)}{S(n)} \tag{4.1}$$

mit

$$W_k = (X_1 + X_2 + \ldots + X_k) - k \cdot \bar{X}(n) \text{ mit } k = 1, 2, 3, \ldots, n \tag{4.2}$$

Hierbei ist $\bar{X}_n$ der arithmetische Mittelwert und $S(n)$ die Standardabweichung der Meßreihe. Hurst beschrieb Zeitreihen mit

$$E[R(n)/S(n)] \sim c_2 \cdot n^H \tag{4.3}$$

wobei c_2 eine Konstante ist und H der Hurst-Parameter der Zeitreihe. Hurst konnte so nachweisen, daß eine Langzeitabhängigkeit existiert, wenn $0,5 < H \leq 1$ gilt. Für $H = 0,5$ kann eine Langzeitabhängigkeit ausgeschlossen werden.
Bei der praktischen Anwendung dieser Methode wird die Meßreihe in K nicht-überlappende Blöcke der Größe n unterteilt und die R/S-Statistik $R(t_i, n)/S(t_i, n)$ in den Blöcken berechnet. t_i entspricht dem Wert in der Meßreihe, mit dem der neue Block beginnt und ergibt sich nach

$$t_i = \frac{(i-1) \cdot N}{K} + 1 \text{ für: } i = 1, \ldots, K.$$

t_1 entspricht dem Wert $X_{t_1} = X_1$ in der Meßreihe. W_k aus Gleichung [4.2] ist hier zu ersetzen durch $W_{t_i+k} - W_{t_i}$.
Zur Bestimmung des Hurst-Parameters werden die Werte der R/S-Statistik in Abhängigkeit von der Blockgröße n aufgetragen. Die Ordinaten- und Abszissenachse werden logarithmisch geteilt. Der Anstieg des Graphen beschreibt den Hurst-Parameter und kann mittels Linearer Regression aus den R/S-Werten gewonnen werden.
Eine Realisierung der R/S-Statistik ist im MATLAB-Skript *rescaled_ bins.m* zu finden. Die detailierte graphische Darstellung bewerkstelligt das Skript *rescaled_ darstellung.m*.

4.2 Variance-Time-Plot

Der Variance-Time-Plot (kurz: *VTP*) gehört zu den am häufigsten verwendeten Schätzverfahren zur Bestimmung des Hurst-Parameters. Unter anderem in (A99), (LTWW94) und (Ros96) wird diese Methode ausführlich vorgestellt.
Ähnlich der R/S-Statistik, wird auch beim Variance-Time-Plot das Verhalten der Meßreihe $(X_t : t = 1, 2, \ldots, n)$ bei verschiedenen Aggregationsstufen untersucht. Beim VTP ist die Meßreihe in k nicht-überlappende Blöcke der Größe m zu unterteilen. Durch Bildung der Mittelwerte in den einzelnen Blöcken ergibt sich eine neue Wertefolge $X^{(m)} = \left(X_k^{(m)} : k = 1, 2, ...\right)$. In den Wertefolgen ist die Varianz $Var(X^{(m)})$ zu ermitteln. Wird die Varianz in Abhängigkeit von der Blockgröße mit jeweils logarithmisch geteilten Achsen dargestellt, kann aus dem Anstieg des resultierenden Graphen der Hurst-Parameter bestimmt werden.
Mittels linearer Regression ist aus dem resultierenden Graphen der Anstieg a zu bestimmen. Durch einsetzen von a in

$$H = 1 - \frac{|a|}{2} \tag{4.4}$$

ergibt sich der Wert für den Hurst-Parameter. Langzeitabhängigkeit liegt vor, wenn $-1 < a < 0$ gilt. Wie bei jeder graphischen Methode sind auch hier die zu betrachtenden Bereiche zu relativieren.
Der Variance-Time-Plot wurde im MATLAB-Skript *varianz_log_opt_bins.m* implementiert. Der Zusatz „log_opt“ beschreibt eine Variante des VTP, in der eine Reduktion der betrachteten Blockgrößen vorgenommen wird. Weitere Informationen zu diesem Thema sind in (GK04) zu finden. Die graphische Darstellung des Variance-Time-Plots ist mit dem Skript *varianz_darstellung.m* durchzuführen.

4.3 Periodogramm

Ausgehend von der diskreten Fouriertransformation, ist das Periodogramm eine Methode zur Abschätzung der spektralen Leistungsdichte. Das Periodogramm ist definiert als:

$$I(\lambda_k) = \frac{1}{N} \cdot \left| \sum_{n=1}^{N} (X_n) \cdot e^{jn\lambda_k} \right|^2 \text{ mit } \lambda_k = \frac{2\pi k}{N}. \tag{4.5}$$

Die Variable k nimmt die Werte $1 \leq k \leq N/2$ an. Darzustellen sind die logarithmierten Werte des Periodogramms gegen die logarithmierten Frequenzen. Das Ergebnis entspricht einer Punktwolke, der mittels linearer Regression eine Gerade zuzuordnen ist. Aus dem Anstieg a der Regressionsgeraden ergibt sich der Hurst-Parameter zu

$$H = -\frac{a-1}{2}. \tag{4.6}$$

Für $-1 < a < 0$ besitzt die untersuchte Meßreihe die Eigenschaft der Selbstähnlichkeit. In der Funktion *periodogramm.m* ist eine mögliche Implementierung des Periodogramms zu finden. Die Darstellung des resulierenden Graphen wurde wiederum getrennt im MATLAB-Skript *periodo_ darstellung.m* implementiert.

Kapitel 5

Analyse

In diesem Kapitel werden die Ergebnisse der Analyse näher vorgestellt. Dabei wird zunächst auf das Meßsystem eingegangen. Anschließend werden die analysierten Meßreihen vorgestellt und auf Abweichungen und Stationarität untersucht.
Verwendung finden die in Kapitel 2 beschriebenen Methoden zur Bestimmung der Verteilungsparameter, der Maximum Likelihood Estimator, der Least Squares Estimator und die Methode der Momente. Betrachtet werden nur die Lognormal-, Pareto- und Weibullverteilung, da hier die aggregierten Zwischankunftszeiten analysiert werden.
Angewandt werden auch die Verfahren zur Schätzung des Hurst-Parameters aus Kapitel 4, um Langzeitabhängigkeit in den Meßreihen nachzuweisen. Um die Güte der Anpassung zwischen empirscher und theoretischer Verteilung (F_X und F_0) zu bestimmen, werden der maximale Abstand d_{max}, der Korrelationskoeffizient r, der Root Mean Square Error (RMS) und der P-P-Plot aus Kapitel 3 verwendet.

5.1 Angaben zum Meßsystem

Es folgt eine kurze Beschreibung des Meßsystems. Ausführliche Betrachtungen sind in (Bru05) und in der noch nicht veröffentlichten Dissertation von Herrn Dipl.-Ing. Kessler (Kes05) zu finden. Für den Aufbau eines Meßsystems sind vor allem die Genauigkeit und die Auflösung des Zeitgebers für die Markierung der Meßdaten, der Datendurchsatz und die Speicherkapazität von Bedeutung.
Werden Messungen nur mit Hilfe einer Netzwerkkarte durchgeführt, werden die Zeitstempel der Pakete zu ungenau vergeben. Der Zeitstempel wird erst erstellt, wenn das Paket vollständig eingetroffen ist und in den Arbeitsspeicher geschrieben wurde. Der Prozess der Erfassung ist nicht deterministisch. Als Gründe hierfür sind unter anderem die IRQ-Verarbeitung, Prioritäten und der Zeitverlust bei einem Kontextwechsel anzugeben.
Weiterhin besteht die Möglichkeit, daß mehrere Frames den gleichen Zeitwert zugeordnet bekommen. Das ist unbedingt zu vermeiden, da gleiche Zeitwerte eine hohe Fehlerquelle in der Analyse des Datenverkehrsaufkommens darstellen. In diesem Zusammenhang wurde eine Netzwerkanalysekarte der Firma eine Endace verwendet. Die Meßkarte DAG 4.2GE führt alle notwendigen Verarbeitungsschritte (mit Hilfe eines Xilinx FPGA) on chip aus. Die ankommenden Frames werden mit einem 64 Bit langen Zeitstempel (in UNIX-Notation) versehen, dessen Auflösung $\frac{1}{2^{32}}$ Sekunden (<100 ns) beträgt. Der Zeitstempel wird bereits generiert und im Meßsystem gespeichert, wenn der Header des Pakets eintrifft.
In Abbildung 5.1 ist wird das Blockschaltbild des Meßsystems dargestellt. Es besteht aus folgenden Kernkomponenten:

- Dual intel Xeon 2400 MHz CPU mit 512 kByte Cache
- Intel E7500 Chipsatz, 2 unabhängige PCI-X Busse, 1 PCI 2.2 Bus
- 2 GByte Dual DDR-RAM Hauptspeicher (im späteren Betrieb Erweiterung auf 3 GByte)
- 10/100 BaseT NIC Intel 82550 für System-Management
- 18 GByte Festplatte (10k UPM SCSI-HDD) für das Betriebssytem
- internes RAID System 3Ware 7500-8 (RAID 0) für die Speicherung der Analysedaten, 1 TB Bruttokapazität
- Netzwerkanalysekarte Endace DAG 4.2GE mit 2-Port GbE-SX Ports
- Debian 3.0 Linux Distibution
- Linux Kernel 2.4.20 (Standardkernel) mit Treibern für Netzanalysekarte

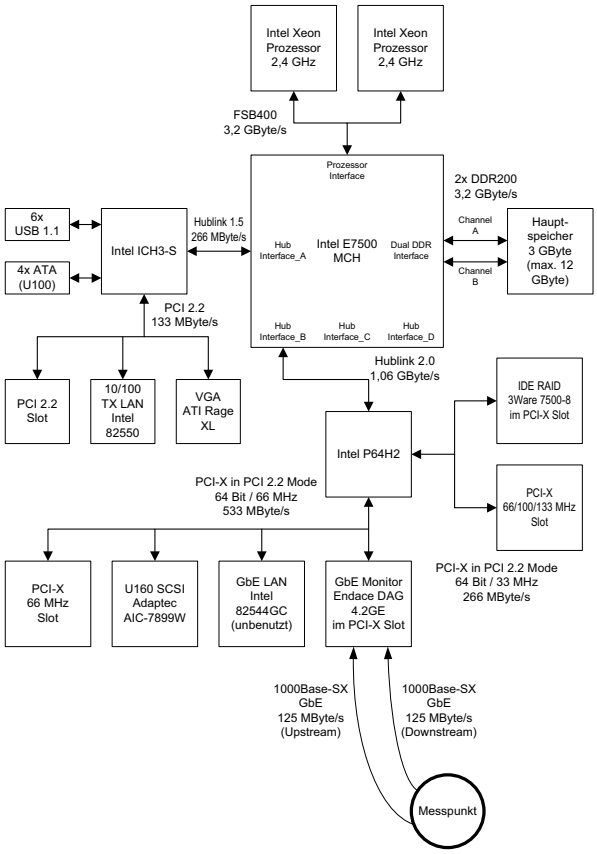


Abbildung 5.1: Blockschaltbild des Gigabit Ethernet Meßsystems

Das hier beschriebene Meßsystem stellt eine Kombination von Hard- und Softwarelösungen dar. So können mit sehr hoher Flexibilität passive Messungen in IP Netzen erfolgen. Eine passive Messung ist eine Messmethode, bei der keine zusätzlichen Meßpaket in den Datenstrom eingebracht werden. Ist die Netzlast zu gering, können mit dieser Meßmethode jedoch nicht genügend Daten gesammelt werden. Aktive und hybride Messungen sind als weitere Methoden zu nennen und werden in (Bru05) eingehend beschrieben. Ebenfalls in (Bru05) und (Kes05) zu finden sind die zugehörigen Performance-Tests.

5.2 Analysierte Datensätze

Gemessen wurde das Datenverkehrsaufkommen an der Universität Rostock. Die Aufzeichung startete am 23.01.2005 um 23:55:11 Uhr und endete am 24.01.2005 um 23:56:14 Uhr. Die Daten standen in einer MySQL-Datenbank (*iptrace_unihro_20050124*) zur Verfügung, mit den Tabellen *flowtab* und *packettab*. In dieser Arbeit waren die Zwischenankunftszeiten und Längen der Pakete zu untersuchen. So wurden die Daten in der Tabelle *packettab* analysiert, mit den in Tabelle 5.1 aufgeführten Feldern.

Feld	Beschreibung
time	auf UNIX basierende Zeitstempel
flags	Angabe der Richtung (0=eingehend, 1=ausgehend)
src_ip	IP-Adresse der Quelle
dst_ip	IP-Adresse des Ziels
ip_proto	Protokoll der Transportschicht
sport	Port der Quelle
dport	Port des Ziels
w_len	Länge des Ethernetrahmens
l2h_len	Länge des Headers auf Schicht 2
l3h_len	Länge des Headers auf Schicht 3
l4h_len	Länge des Headers auf Schicht 4

Tabelle 5.1: Übersicht der Felder der Tabelle *packettab*

Für die Betrachtung der Zwischenankunftszeiten waren nur die Zeitstempel im Feld **time** von Interesse. Bei der Analyse der Paketlängen wurde der Payload auf der Vermittlungsschicht (Schicht 3 des ISO/OSI-Referenzmodells) betrachtet. Hierzu waren die Längen der Header der Schichten 2 und 3 von den Werten im Feld **w_len** abzuziehen.
Während der Analyse wurden der eingehende und der ausgehende Netzwerkverkehr getrennt betrachtet. Dabei wurden in 725.048.123 Paketen $3{,}19224 \cdot 10^{11}$ Byte empfangen bzw. 542.402.418 Pakete in $1{,}63843 \cdot 10^{11}$ Byte gesendet. Das hier angegebene Datenvolumen bezieht sich auf den Payload auf der Vermittlungsschicht.
Aus Gründen der Datenreduktion wurden die gesendeten bzw. empfangenen Pakete innerhalb einer jeden Sekunde gezählt. In den folgenden Abschnitten wird dieses Vorgehen als Bildung von Bins[1] bezeichnet, hier mit einem Intervall von 1s. Tabelle 5.2 zeigt einige Werte der deskritiven Statistik für die Zwischenankunftszeiten. Für die Paketlängen, wurden

	empfangen	gesendet
arithmetisches Mittel	8.385,646	6.273,232
Median	7.293	6.173
Standardabweichung	4.714,546	1.304,753
Minimalwert	3.811	3.517
Maximalwert	3.1264	13.815

Tabelle 5.2: Eigenschaften der aggregierten Zwischenankunftszeiten [Pakete/Sekunde]

die innerhalb des Intervalls übertragenen bzw. gesendeten Byte addiert. Die Werte der deskriptiven Statistik für die summierten Paketlängen sind in Tabelle 5.3 zusammenfassend dargestellt. Wie zu erwarten, war beim eingehenden Verkehr ein höheres Datenaufkommen zu verzeichnen. Inwiefern Abweichungen hier eine Rolle spielen, wird in Abschnitt 5.3 erläutert. Detaillierte Darstellungen der Verläufe sind ebenfalls in diesem Abschnitt zu finden.

1 bin - deutsch:(*Behälter*)

	empfangen	gesendet
arithmetisches Mittel	$3{,}69203 \cdot 10^6$	$1{,}89494 \cdot 10^6$
Median	3.772.672	1.976.682
Standardabweichung	$8{,}09843 \cdot 10^5$	$4{,}45196 \cdot 10^5$
Minimalwert	1.587.619	1.132.823
Maximalwert	12.457.107	3.563.287

Tabelle 5.3: Eigenschaften der summierten Paketlängen [Byte/Sekunde]

5.3 Abweichungen in den Meßreihen

In diesem Abschnitt werden die zu analysierenden Datensätze in Hinblick auf Abweichungen untersucht. Abweichungen (auch: „Ausreißer“) sind Werte, die nicht in eine erwartete Meßreihe passen und die Ergebnisse der Analyse negativ beeinflußen könnten.
Hierzu werden in Abbildung 5.2 die aggregierten Zwischenankunftszeiten über die gesamte Messdauer dargestellt. Gebildet wurden die Bins mit der Bingröße 1s. Auffällig

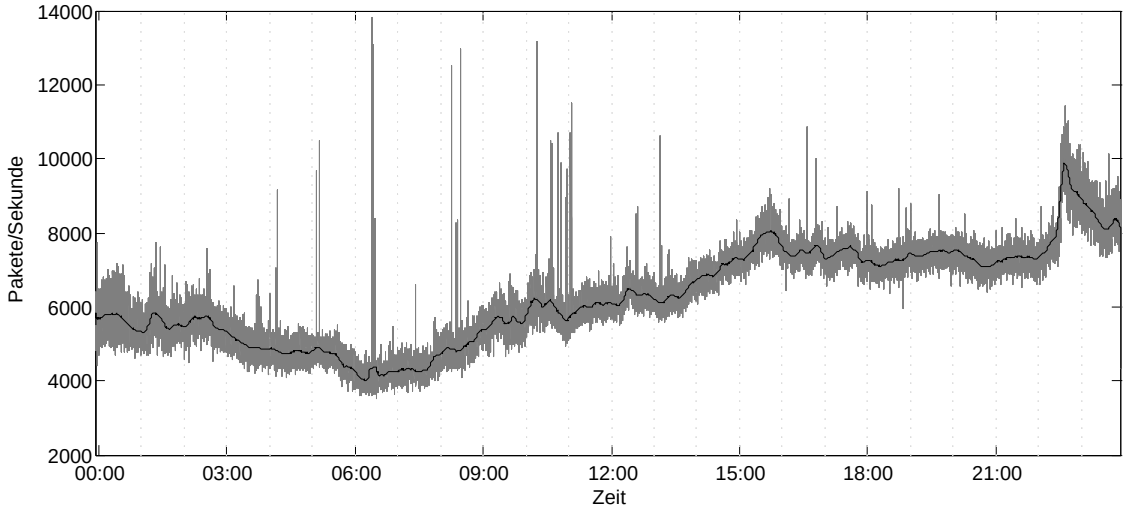


Abbildung 5.2: Gesendete Pakete in Abhängigkeit von der Zeit

sind die vereinzelt auftretenden Spitzen zwischen 4 und 12 Uhr. Grund war jeweils eine schnelle Abfolge von Paketen, die größtenteils nur Headerdaten enthielten. Vergleichend hierzu wird in Abbildung 5.3 das übertragene Datenvolumen dargestellt. Die schwarz gekennzeichneten Linien zeigen den mittels gleitenden Mittelwert geglätteten Verlauf. Dies ermöglicht eine bessere Bewertung. Es ist zu erkennen, daß die oben erwähnten Spitzen nur einen geringen Einfluß auf das Datenvolumen haben.
Ebenfalls auffällig in Abbildung 5.2 ist der zwischen 22 und 23 Uhr auftretende Versatz. In Abbildung 5.3 ist zu erkennen, daß beim Verlauf des Datenvolumens kein solcher Versatz auftritt. Es konnte bis zur Abgabe dieser Arbeit nicht geklärt werden, ob zu diesem Zeitpunkt besondere Vorkommnisse zu verzeichnen waren. An dieser Stelle sei auf weitere

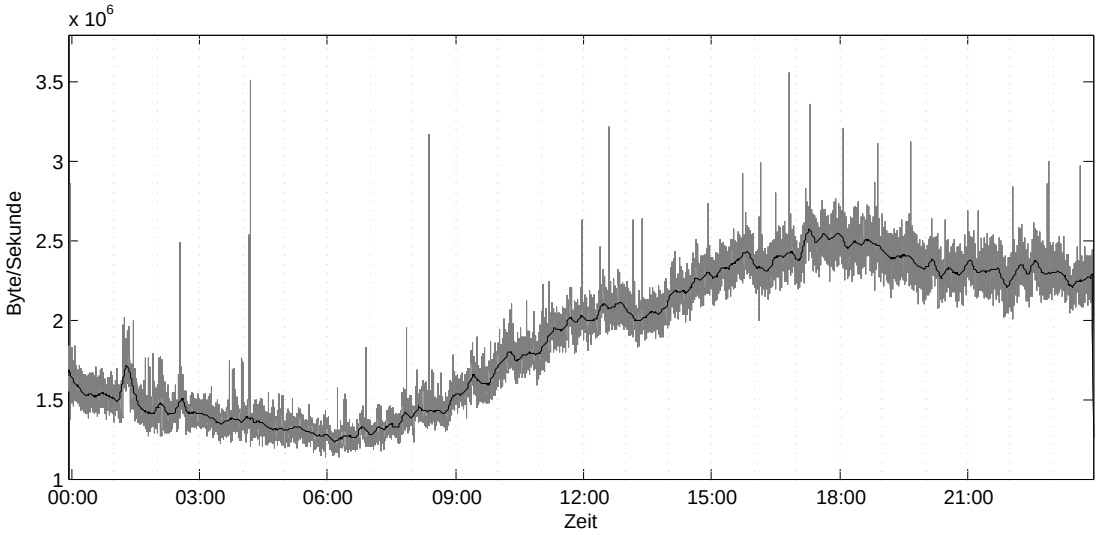


Abbildung 5.3: Paketlängen des gesendeten Verkehrs in Abhängigkeit von der Zeit

Spekulationen verzichtet. Für den eingehenden Verkehr ist in Abbildung 5.4 die Anzahl der empfangenen Pakete pro Sekunde dargestellt. Zur Wahrung der Übersichtlichkeit wird die

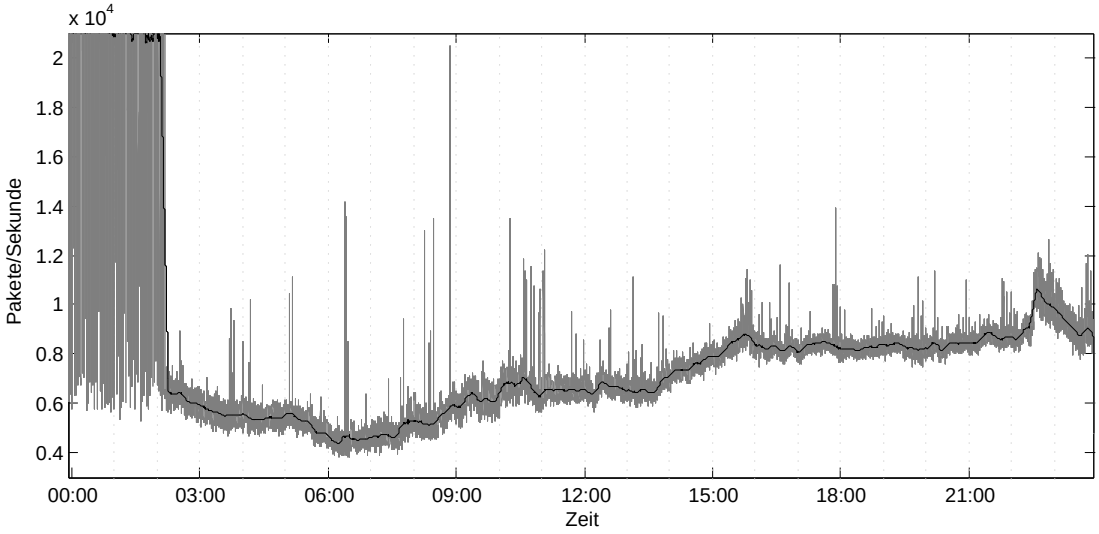


Abbildung 5.4: Empfangene Pakete in Abhängigkeit von der Zeit

Darstellung auf 21000 Pakete pro Sekunde begrenzt, obwohl der Maximalwert bei 31264 Paketen pro Sekunde lag. Die vollständige Darstellung ist in Anhang A zu finden (vgl. Abbildung A.1). Wie bereits beim ausgehenden Verkehr treten auch beim eingehenden Verkehr die bereits erwähnten Spitzen und der Versatz zwischen 22 und 23 Uhr auf. Eine weitere Besonderheit ist der Bereich zwischen Beginn der Messung und etwa 2:10 Uhr. Hier sind deutliche Abweichungen vom sonstigen Datenverkehrsaufkommen zu erkennen. Zu erklären ist dies ebenfalls durch schnelle Abfolgen von sehr kleinen Paketen, die nur Headerdaten enthielten. Untermauert wird diese Aussage durch den Verlauf des empfan-

genen Datenvolumens, wie in Abbildung 5.5 zu sehen. Das Datenvolumen wird in diesem Bereich nicht signifikant beeinflußt. Aufgrund der Abweichungen ist es nicht sinnvoll, eine Analyse des gesamten Bereichs durchzuführen. Sollen 24 Stunden untersucht werden, ist entweder ein zweiter Datensatz erforderlich oder eine Einschränkung des Betrachtungsbereichs. In Zusammenhang mit dieser Arbeit sei hierauf jedoch verzichtet.
Die Darstellungen zeigen bereits, daß die Meßreihe nicht stationär ist. Dies wird im nächsten Abschnitt genauer untersucht.

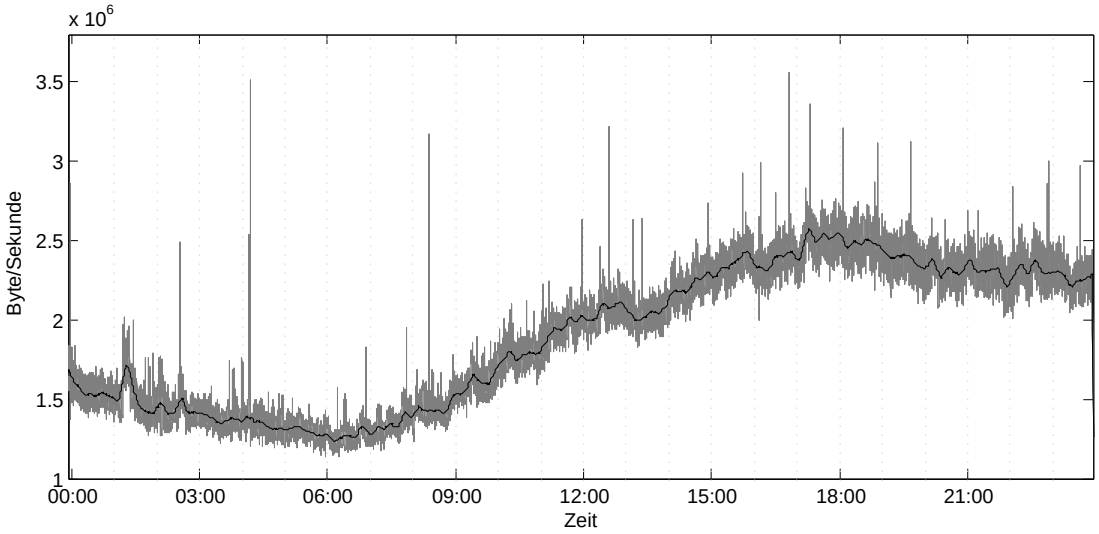


Abbildung 5.5: Paketlängen des empfangenen Verkehrs in Abhängigkeit von der Zeit

5.4 Untersuchung auf Stationarität

Der Begriff der Stationarität wurde bereits in Abschnitt 1.2.4 erläutert. Dieser Abschnitt befasst sich mit dem Test der zu analysierenden Datensätze auf Stationarität. Das Problem der Stationarität gewinnt mit der Vergrößerung des Betrachtungszeitraums von Bedeutung. Bei relativ kurzen Betrachtungen von einigen Minuten bzw. einer Stunde kann von einer Stationarität ausgegangen werden. Nicht-stationäre Anteile sind zu beachten, wenn der Beobachtungszeitraum mehrere Stunden überschreitet, wie in (Fel01) und (Bru05) zu sehen. Dies verdeutlichen ebenfalls die Abbildungen 5.2 bis 5.5.
Hier sind über den Tagesverlauf große Schwankungen zu erkennen, die unterschiedliche Belastungssituationen beschreiben. Vergleichend werden in Tabelle 5.4 die arithmetischen Mittelwerte und Standardabweichungen von zwei ausgewählten Bereichen (Dauer: 1h) dargestellt. Mit dem arithmetischem Mittel als Wert für das erste und der Standardabwei-

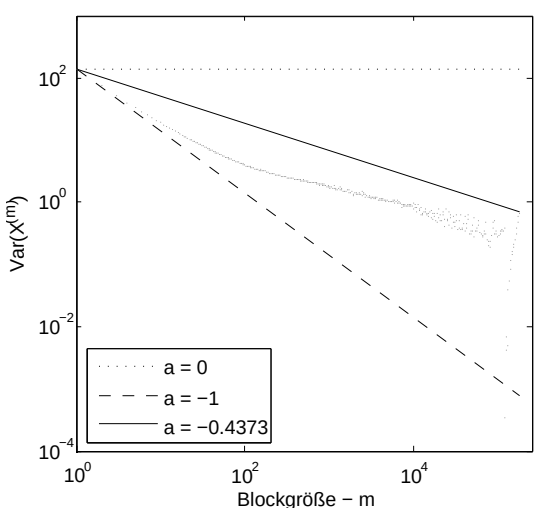
		7 Uhr bis 8 Uhr		18 Uhr bis 19 Uhr	
		empfangen	gesendet	empfangen	gesendet
arithmetisches Mittel	[Pakete/s]	4.839,14	4.376,93	8.237,72	7.220,27
Standardabweichung	[Pakete/s]	466,14	265,49	199,51	238,04
arithmetisches Mittel	[Byte/s]	$2{,}22013 \cdot 10^6$	$1{,}34157 \cdot 10^6$	$4{,}44025 \cdot 10^6$	$2{,}48616 \cdot 10^6$
Standardabweichung	[Byte/s]	$1{,}83274 \cdot 10^5$	$6{,}56242 \cdot 10^4$	$1{,}79877 \cdot 10^5$	$7{,}89347 \cdot 10^4$

Tabelle 5.4: Vergleich der Mittelwerte und Standardabweichungen

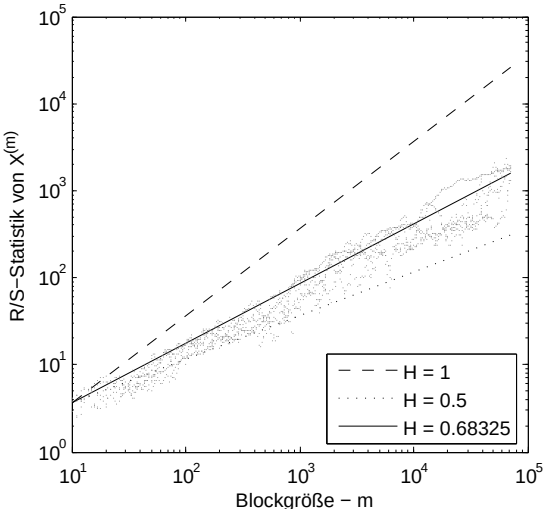
chung als Wert für das zweite Moment entspricht dies dem Verfahren der Fensterbildung (vgl. Abschnitt 1.2.4). Zu erkennen sind deutliche Unterschiede zwischen den Erwartungswerten. Der Unterschied der Standardabweichungen fällt weniger ins Gewicht.
Schlußfolgernd kann bei der Betrachtung über die gesamte Meßdauer nicht von Stationarität ausgegangen werden. In den folgenden Abschnitten werden somit verschiedene Belastungssituationen analysiert. Untersucht wird der Bereich 7 bis 8 Uhr als Niedriglastsituation und der Bereich zwischen 18 und 19 Uhr als Hochlastsituation, jeweils das niedrigste bzw. höchste Netzwerkverkehrsaufkommen betrachtend.

5.5 Bestimmung der Hurst-Parameter

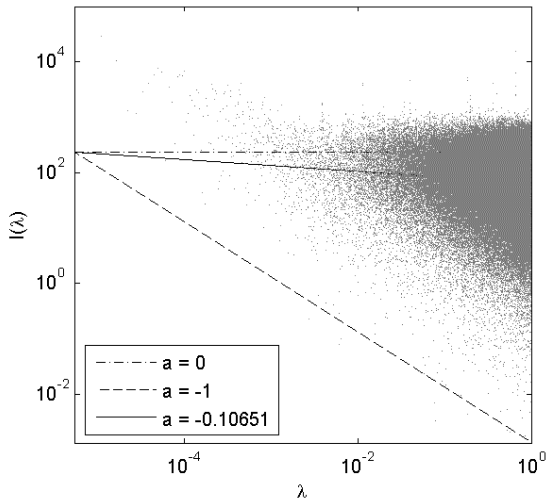
In diesem Abschnitt wird auf die ermittelten Hurst-Parameter eingangen. Dazu werden die in Kapitel 4 beschriebenen Methoden angewandt. In Abbildung 5.6(a) wird der VTP und in Abbildung 5.6(b) die R/S-Statistik exemplarisch für den eingehenden Verkehr zwischen 18 und 19 Uhr dargestellt. Verwendet wurde hier eine Bingröße von 10ms und so war der



(a) *Variance-Time-Plot*



(b) *R/S-Statistik*



(c) *Periodogramm*

Abbildung 5.6: Darstellung der Methoden zur Bestimmung des Hurst-Parameters

Hurst-Parameter innerhalb der 360.000 Bins zu bestimmen. Der ermittelte Anstieg des VTP beträgt a=-0.4373. Das entspricht nach Gleichung [4.4] einem Hurst-Parameter von H=0,78135. Der Hurst-Parameter aus der R/S-Statistik ergab sich zu H=0.68325 direkt aus dem Anstieg der Regressionsgeraden.
Vergleichend dazu ist in Abbildung 5.6(c) das Periodogramm dargestellt, dessen Anstieg

mit a=-0.10651 ermittelt werden konnte. Durch Einsetzen des Anstiegswerts in Gleichung [4.6] beträgt der Wert für den Hurstparameter H=0,5533. In Tabelle 5.5 werden die Hurstparameter der in dieser Arbeit betrachteten aggregierten Zwischenankunftszeiten zusammenfassend dargestellt. Auf die zugehörigen Abbildungen sei hier verzichtet. Der Tabelle ist zu entnehmen, daß die Werte für den Hurstparameter jeweils zwischen

	07 - 08 Uhr			
	H_{VTP}	$H_{R/S}$	H_{Peri}	H
empfangsseitig	0,8494	0,7238	0,7039	0,7590
sendeseitig	0,8061	0,6550	0,6159	0,6923
	18 - 19 Uhr			
empfangsseitig	0,7813	0,6833	0,5533	0,6726
sendeseitig	0,7843	0,6530	0,5934	0,6769

Tabelle 5.5: Zusammenfassung der geschätzten Hurstparameter

$0,5 < H < 1$ liegen. Das legt den Schluß nahe, daß es sich hierbei um selbstähnliche Ankunftsprozesse handelt. Ebenfalls zu erkennen ist, daß die geschätzten Hurstparameter stark schwanken, je nach verwendetem Schätzverfahren. Dieses Problem der genauen Bestimmung wird auch in (KFR02) gezeigt.

5.6 Analyse verschiedener Belastungssituationen

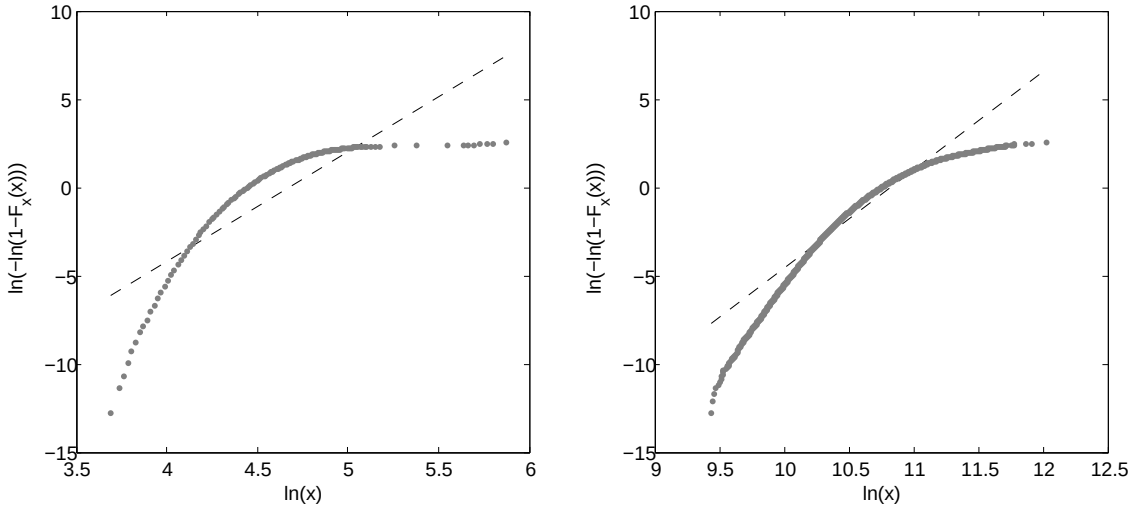
Wie in Abschnitt 5.4 gezeigt, kann nicht von Stationarität ausgegangen werden, wenn die gesamte Meßdauer betrachtet wird. In den folgenden Abschnitten werden einige ausgewählte Zeitbereiche näher betrachet. Dabei wird unterschieden zwischen Hochlast- und Niedriglastsituationen. Anwendung finden die in Kapitel 2 vorgestellten Verfahren zur Schätzung der Verteilungsparameter. Aus Gründen der Datenreduktion wurden wiederum Bins gebildet, hier mit einem Intervall (auch: *Bingröße*) von 10ms. Im nächsten Abschnitt wird die Anwendung der Verfahren zur Schätzung der Verteilungsparameter exemplarisch für einen ausgewählten Datensatz dargestellt.

5.6.1 Anwendung der Verfahren zur Gewinnung der Verteilungsparameter

In den folgenden Abschnitten werden die Schätzverfahren aus Kapitel 2 angewandt und die Ergebnisse detailliert dargestellt. Verwendet werden die Zwischenankunftszeiten und Paketlängen des eingehenden Verkehrs zwischen 18 und 19 Uhr vom 24.01.2005, bei einer Bingröße von 10 ms. Es ist zu erwarten, daß die Methode der kleinsten Quadrate nur ungenaue Werte der Parameter liefern wird. Genauere Werte werden der Momentenschätzer und der MLE ergeben. Die größten Anpassungen werden, entsprechend der Abbildung 5.16, für die Lognormal- und Weibullverteilung erwartet. Die Exponentialverteilung wird nicht in die Analyse einbezogen, da hier ein aggregierter Ankunftsprozess vorliegt.

5.6.1.1 Verfahren der Weibullverteilung

Beginnend mit den Verfahren zur Schätzung der Weibullparameter werden in Abbildung 5.7 die „probability plots“ der aggregierten Zwischenankunftszeiten und summierten Paketlängen dargestellt. Es ist zu erkennen, daß in beiden Verläufen eine deutliche Verände-



(a) ***aggregierte Zwischenankunftszeiten*** **(b)** ***summierte Paketlängen***

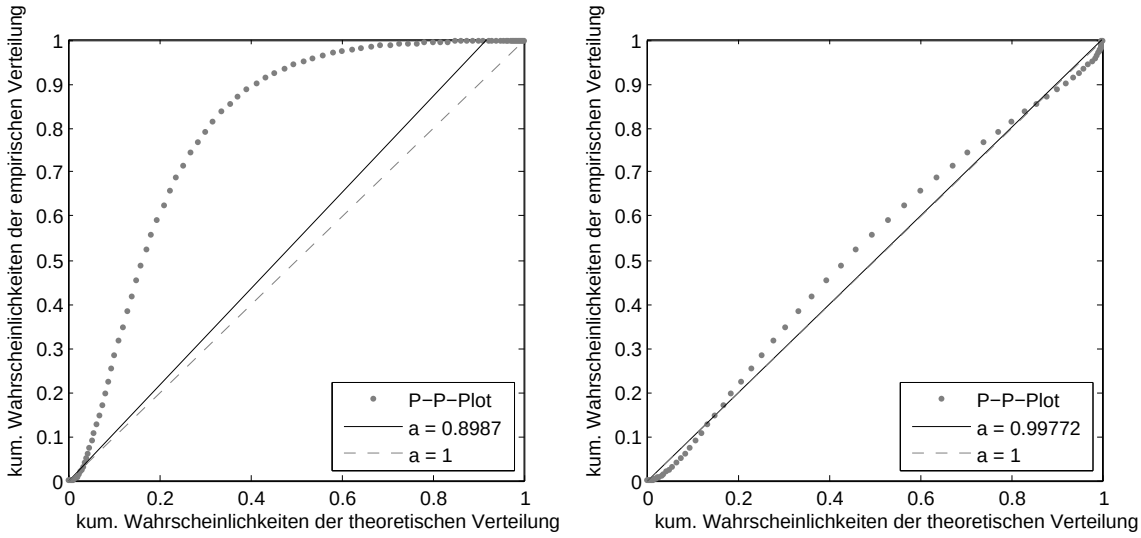
Abbildung 5.7: LSE der Weibullverteilung

rung des Anstiegs auftritt. Obwohl dieser „Bogen“ bei der Darstellung für die Paketlängen deutlich schwächer ausgeprägt ist, deuten die Verläufe darauf hin, daß eine Weibullverteilung kein geeignetes Modell ist. Die geschätzten Werte der Parameter sind der Tabelle 5.6 zu entnehmen. Hierbei sind d_{max} die maximale Differenz, r der Korrelationskoeffizient und RMS der Root Mean Square Error zwischen empirischer und theoretischer Verteilung.
Der Vergleich der geschätzten Parameter zeigt teilweise deutliche Abweichungen, besonders bei den Zwischenankunftszeiten. Interessant hierbei ist, daß der Momentenschätzer

	Ankunftszeiten			Paketlängen		
	Momente	MLE	LSE	Momente	MLE	LSE
$\hat{\alpha}_w$	8,3815	5,7895	6,1881	5,0040	4,3941	5,5237
$\hat{\beta}$	86,8784	87,0190	107,3836	48.357,08	48.422,73	50.231,82
d_{max}	0,0643	0,1071	0,5008	0,0550	0,0557	0,1369
r	0,99866	0,99675	0,90660	0,99725	0,99715	0,98965
RMS	0,0211	0,0436	0,2214	0,0278	0,0342	0,0687

Tabelle 5.6: Vergleich der Schätzungen, Weibullverteilung

scheinbar bessere Werte für die Parameter liefert als der Maximum Likelihood Estimator. Der Tabelle ist ebenfalls zu entnehmen, daß die Methode der kleinsten Quadrate nicht dazu geeignet ist, um genaue Werte für die Parameter der Weibullverteilung zu bestimmen. Vergleichend hierzu sind in Abbildung 5.8 die P-P-Plots für die Werte aus der Least Squares Schätzung und Momentenschätzung dargestellt. Der eigentliche Verlauf des P-

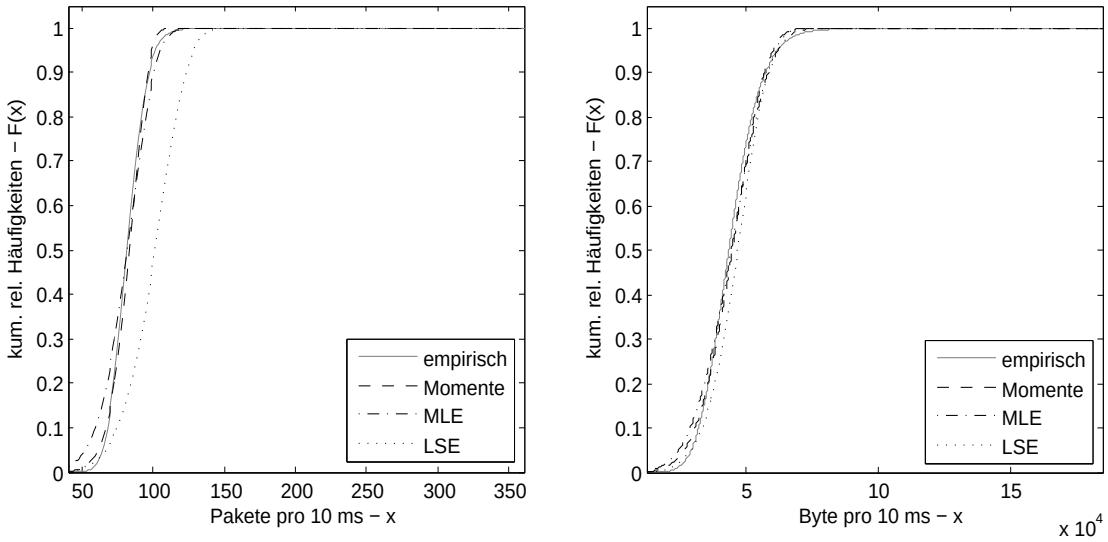


(a) ***Least Squares Estimator*** **(b)** ***Momentenschätzer***

Abbildung 5.8: P-P-Plots der Schätzungen für die Weibullverteilung

P-Plots wird durch die Punkte dargestellt. Weiterhin werden zwei Geraden gezeigt. Hier entspricht die durchgezogene Linie dem mittels linearer Regression ermittelten Abstieg. Der ideale Verlauf mit dem Anstieg $a = 1$ (vgl. Abschnitt 3.5) wird durch die gestrichelte Linie vergleichend dargestellt.

Den Abbildungen ist zu entnehmen, daß die Vermutung der unzureichenden Genauigkeit des LSE bestätigt wird. Die Methode der Momente liefert jedoch Werte, die auf eine mögliche Weibullverteilung der Zwischenankunftszeiten und Paketlängen schließen lassen. Es ist zu überprüfen, ob nicht eine andere Verteilung einen höheren Grad der Anpassung ermöglicht. Abschließend werden in Abbildung 5.9 die theoretischen und empirschen Verteilungsfunktionen der Zwischenankunftszeiten und der Paketlängen dargestellt.

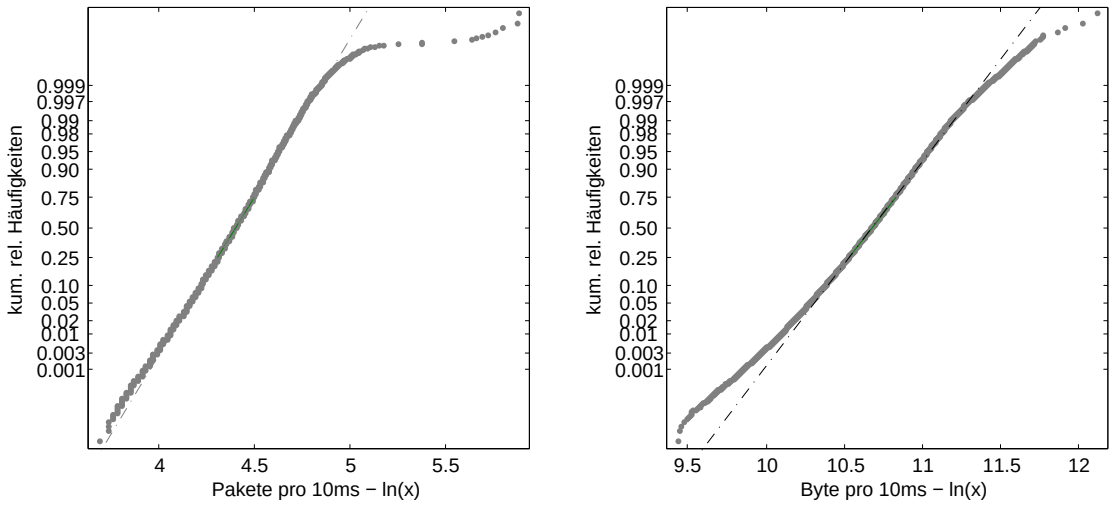


(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.9: Vergleich der geschätzten Verteilungsparameter (Weibull)

5.6.1.2 Verfahren der Lognormalverteilung

Aufgrund des in Abschnitt (1.3.2) gezeigten Zusammenhangs zwischen Normalverteilung und Lognormalverteilung kann mittels „Normal Probability Plot“ geprüft werden, ob die Zwischenankunftszeiten und Paketlängen lognormalverteilt sind (Abbildung 5.10). Wie



(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.10: Normal Probability Plot

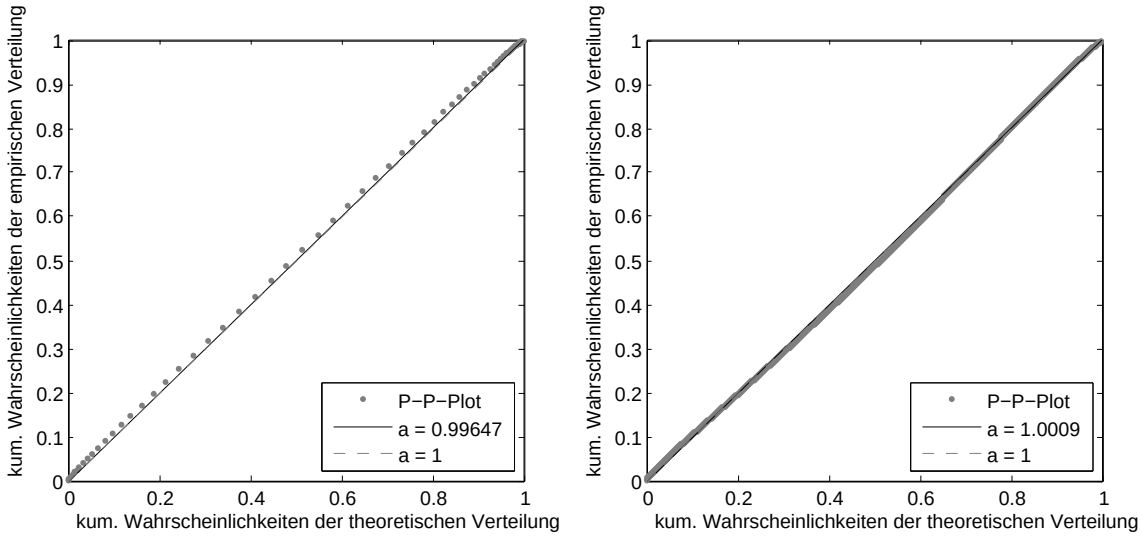
in den Darstellungen zu sehen, entsprechen die Verläufe in großen Teilen einer Geraden. Lediglich die Schwänze der Verteilungen weichen von der Geraden ab. Hier ist eine höhere Anpassung an die empirische Verteilungsfunktion als bei der Weibullverteilung zu erwarten. Die Parameter wurden aus den Darstellungen nicht geschätzt. Die mittels Momen-

tenschätzer und MLE bestimmten Parameter, sind der Tabelle 5.7 zu entnehmen. Hierbei

	Ankunftszeiten		Paketlängen	
	Momente	MLE	Momente	MLE
$\hat{\mu}_L$	4,4014	4,4014	10,6755	10,6751
$\hat{\sigma}_L$	0,1407	0,1412	0,2260	0,2294
d_{max}	0,0126	0,0133	0,0135	0,0148
r	0,99994520	0,99994265	0,99984939	0,99981806
RMS	0,00586	0,00590	0,00643	0,00744

Tabelle 5.7: Vergleich der Schätzungen, Lognormalverteilung

sind wiederum d_{max} die maximale Differenz, r der Korrelationskoeffizient und RMS der Root Mean Square Error. Die Tabelle zeigt, daß auch hier der Momentenschätzer eine höhere Anpassung an die empirische Verteilungsfunktion ermöglicht. Die Abweichungen der geschätzten Parameter fallen hier sehr gering aus. Auch die verwendeten Gütekriterien zeigen nur geringe Unterschiede. Die Vermutung, daß die in MATLAB gegebene Funktion *lognfit* nur ungenaue MLE-Werte liefert, bestätigte sich nicht. Eine eigene Implementierung des MLE führte zu vergleichbaren Ergebnissen.
Der direkte Vergleich mit Tabelle 5.6 legt den Schluß nahe, daß die aggregierten Zwischenankunftszeiten und summierten Paketlängen für den betrachten Zeitbereich lognormalverteilt sind. In Abbildung 5.11 sind die P-P-Plots der Momentenschätzer dargestellt. Die Darstellungen zeigen jeweils eine Gerade mit Anstieg $m \approx 1$. Krümmungen wie bei

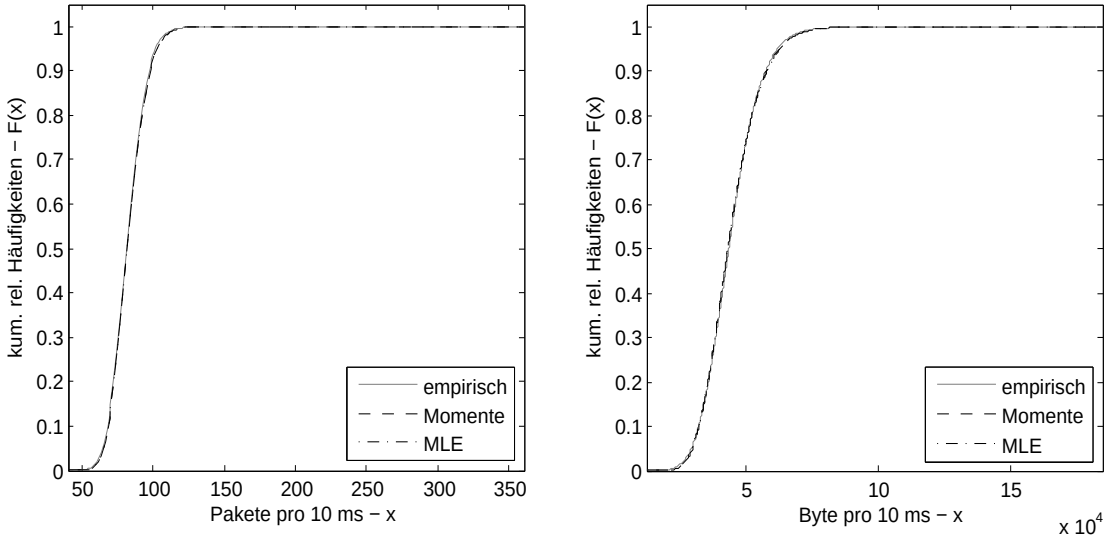


(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.11: P-P-Plots der Momentenschätzer (Lognormal)

der Weibullverteilung sind hier nicht zu erkennen. Aufgrund der Form wird für die Paretoverteilung ein geringerer Grad der Anpassung erwartet. Abschließend werden in Abbildung 5.12 noch einmal die theoretischen und empirischen Verteilungsfunktionen vergleichend dargestellt. Unterschiede zwischen der empirischen und den theoretischen Verteilungen

sind fast nicht existent.



(a) ***aggregierte Zwischenankunftszeiten*** **(b)** ***summierte Paketlängen***

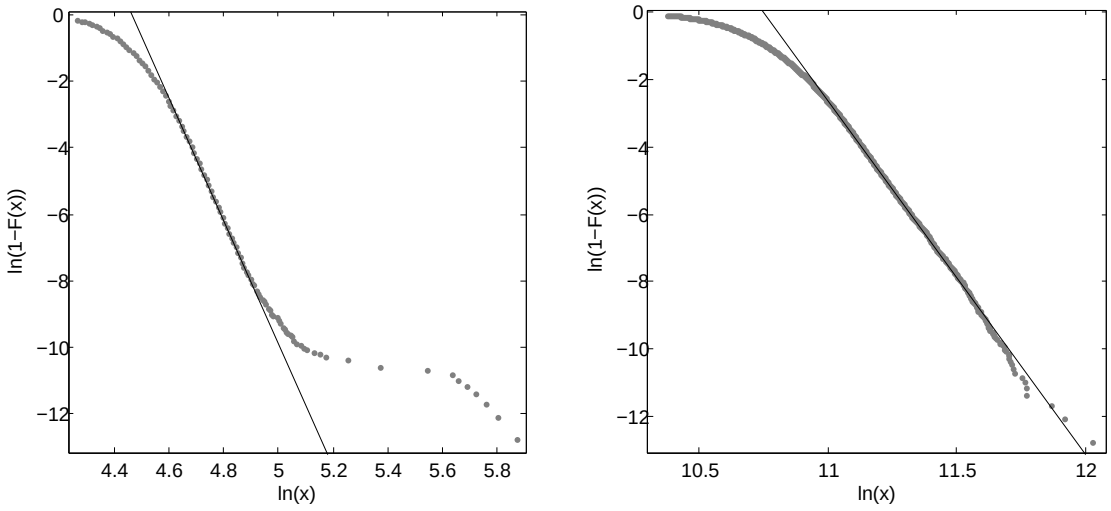
Abbildung 5.12: Vergleich der geschätzten Verteilungsparameter (Lognormal)

5.6.1.3 Verfahren der Paretoverteilung

Bereits in Abschnitt (1.3.3) wurde die Paretoverteilung ausführlich vorgestellt. Die Dichtefunktion der Paretoverteilung ist in allen Bereichen monoton fallend (fallender Monotoniebogen). Die hier betrachteten Folgen weisen jedoch nicht ein solches Verhalten auf. Für eine mögliche Beschreibung mittels Paretoverteilung war der untere Teil der Verteilungsfunktion von der Betrachtung auszuschließen. Der Nachteil liegt hier in der Veränderung der zu beschreibenden Messreihe. Als Kompromiß wurden die oberen 85% der Verteilung betrachtet, so daß die Messreihe nur gering verändert wird.

Die Darstellung des LSE (vgl. Abb. 5.13) der summierten Paketgrößen zeigt über weite Bereiche einen linearen Verlauf. Eine Paretoverteilung ist für die oberen 85% der summierten Paketlängen nicht auszuschließen. Für die aggregierten Zwischenankunftszeiten ist das Ergebnis nicht eindeutig. Es ist ein linearer Bereich zu erkennen, jedoch mit Abweichungen an den Schwänzen, die nicht zu vernachlässigen sind. Die geschätzten Parameter und die Maße für den Grad der Anpassung werden in Tabelle 5.8 zusammengefasst.

Aufgrund der vorgenommenen Veränderung der empirischen Verteilungsfunktion war damit zu rechnen, daß im unteren Bereich der Verteilung der Grad der Anpassung abnimmt. So wurden die maximale Distanz d_{max}, der Korrelationskoeffizient r und der Root Mean Square Error RMS nur für die oberen 50% der Verteilungen ermittelt. Der Tabelle ist zu entnehmen, daß der Momentenschätzer Werte für die Parameter der Paretoverteilung liefert, die eine bessere Anpassung an die empirische Verteilungsfunktion erlauben. Für die



(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.13: LSE der Paretoverteilung

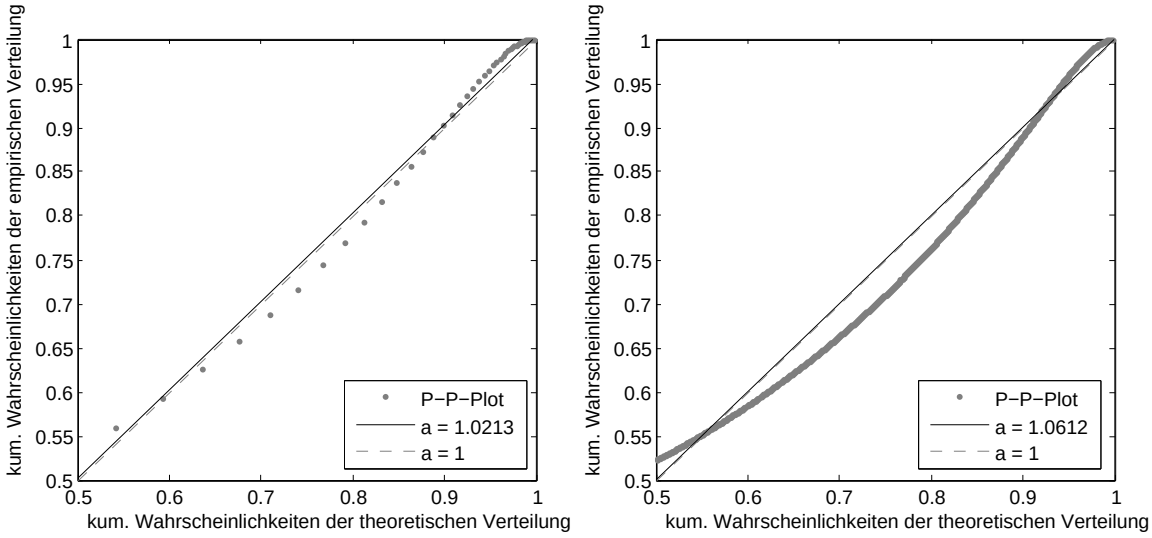
	Ankunftszeiten			Paketlängen		
	Momente	MLE	LSE	Momente	MLE	LSE
$\hat{\alpha}_p$	9,7501	5,6216	18,3070	6,4257	3,3466	10,4744
$\hat{t}_0$	76,6011	71	86,6391	39.615,90	34.240,00	46.632,32
d_{max}	0,0251	0,0830	0,2674	0,0418	0,0869	0,2299
r	0,9964	0,9635	0,9802	0,9935	0,9898	0,9837
RMS	0,00992	0,0435	0,0433	0,0210	0,0643	0,0588

Tabelle 5.8: Vergleich der Schätzungen, Paretoverteilung

aggregierten Zwischenankunftszeiten liefert der Maximum Likelihood Estimator Werte, die nur eine geringe Anpassung ermöglichen. Grund hierfür ist die Definition des MLE für den Parameter t_0 (vgl. Abschnitt 2.4.3). Sowohl α_p als auch t_0 haben starken Einfluß auf die Form der Paretoverteilung. Ist t_0 fest vorgegeben, kann eine Anpassung nur noch über den Parameter α_p erfolgen.

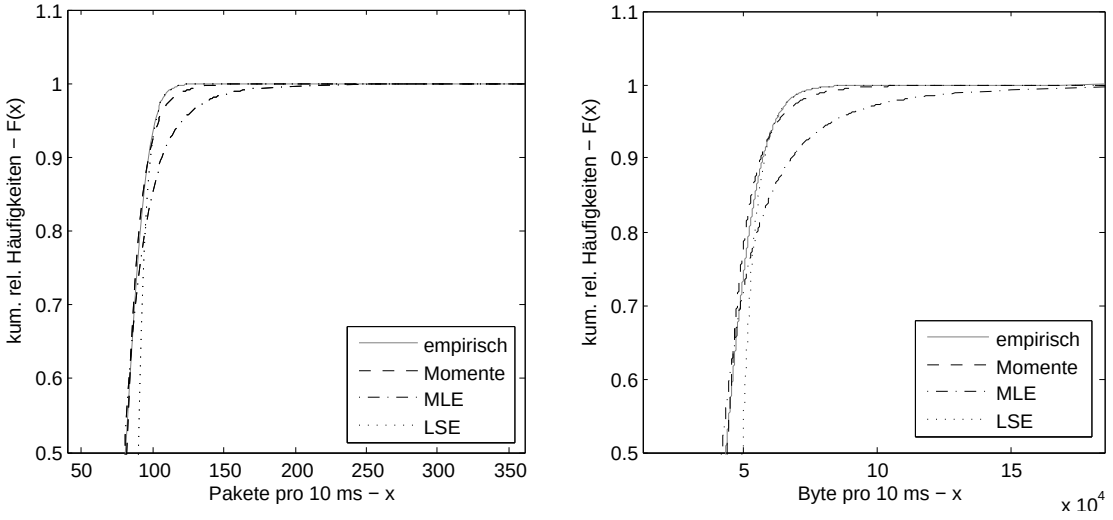
In Abbildung 5.14 sind die P-P-Plots der Momentenschätzer dargestellt. Wie bei der Weibullverteilung treten auch hier Krümmungen um die Regressionsgerade auf. Betrachtet wurden für die P-P-Plots nur die oberen 50% der empirischen und theoretischen Verteilungen. Eine mögliche Paretoverteilung der summierten Paketlängen und aggregierten Zwischenankunftszeiten ist im oberen Bereich nicht auszuschließen.

Abschließend zu diesem Abschnitt werden die oberen 50% der theoretischen und empirischen Verteilungen noch einmal vergleichend dargestellt (Abb. 5.15). Die Darstellung für den Gesamtbereich ist in Abbildung A.2 im Anhang zu finden.



(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.14: P-P-Plots der Momentenschätzer (Pareto)



(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.15: Vergleich der geschätzten Verteilungsparameter (Pareto)

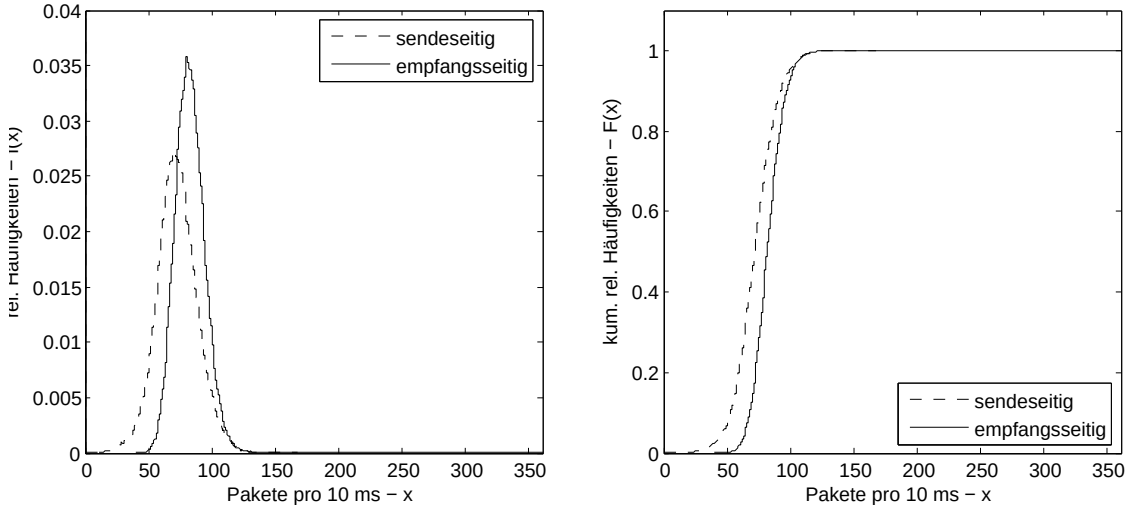
5.6.2 Hochlastsituation

Betrachtet wird in diesem Abschnitt wiederum der Zeitbereich zwischen 18 und 19 Uhr vom 24.01.2005. Innerhalb dieses Zeitraums wurden in 29.655.788 Paketen $1{,}59848 \cdot 10^{10}$ Byte empfangen. Gesendet wurden $8{,}95019 \cdot 10^{9}$ Byte in 25.992.983 Paketen. Die hier angegebenen Paketlängen beziehen sich auf den Payload der Vermittlungsschicht. In Tabelle 5.9 sind einige Eigenschaften des hier betrachteten Bereichs zusammengefasst. Die

	Ankunftszeiten [μs]		Paketlängen [Byte]	
	empfangen	gesendet	empfangen	gesendet
arithmetisches Mittel	121,393	138,499	539,014	344,331
Median	84,490	84,892	113	32
Standardabweichung	130,785	189,659	591,549	520,401
Minimalwert	0,656	0,656	8	8
Maximalwert	2.318,31	18.586,70	1.480	1.480

Tabelle 5.9: Eigenschaften der Zwischenankunftszeiten und Paketlängen (Hochlast)

maximalen Paketlängen entsprechen der Maximallänge der hier aufgetretenen Ethernetrahmen (1518 Byte) nach Abzug der Header von Schicht 2 und Schicht 3 im ISO/OSI-Referenzmodell. Für die Zwischenankunftszeiten sind in Abbildung 5.16 die zugehörigen empirischen Dichte- und Verteilungsfunktionen dargestellt. Die Dichte- und Verteilungs-

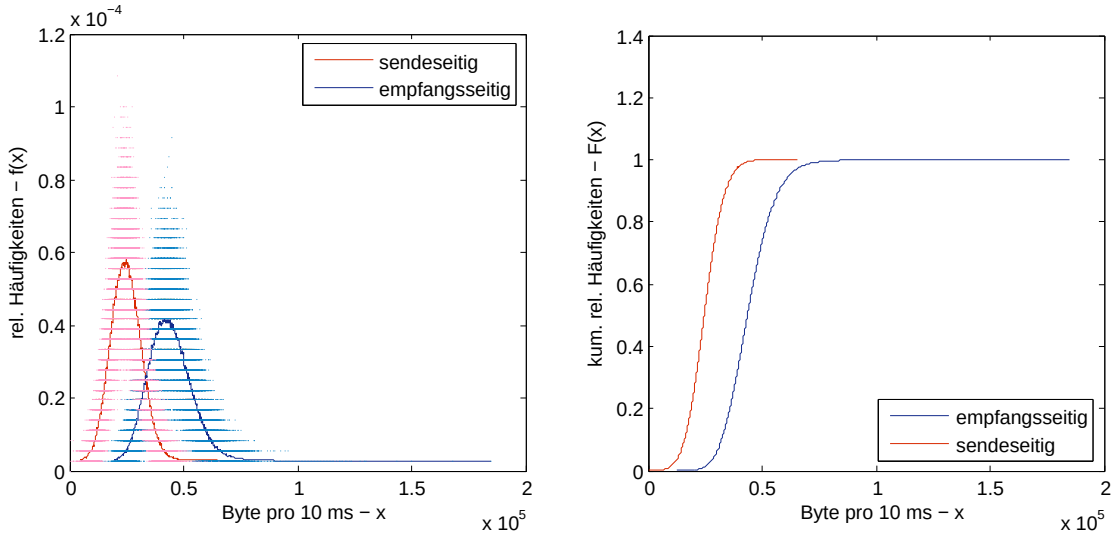


(a) *Verteilungsdichten* **(b) *Verteilungsfunktionen***

Abbildung 5.16: empirische Verteilungen der agg. Zwischenankunfszeiten (18 bis 19 Uhr)

funktionen der summierten Paketlängen sind in Abbildung 5.17 zu sehen. Auffallend sind die Dichtefunktionen der Paketlängen. Die Form ist nicht klar definiert. Mit Hilfe eines gleitenden Mittelwerts wurden die Dichtefunktionen geglättet, um den Verlauf zu verdeutlichen. Es ist klar, daß die Dichtefunktionen der Paketlängen ungeeignet sind, um einen

Vergleich anzustellen. Die hier gewählten Kriterien für den Grad der Anpassung beziehen sich auf die Verteilungsfunktionen. 5.17 dargestellt.



(a) ***Verteilungsdichten*** **(b)** ***Verteilungsfunktionen***

Abbildung 5.17: empirische Verteilungen der summierten Paketlängen (18 bis 19 Uhr)

5.6.2.1 Eingehender Verkehr

Die Analyse der Verteilungsfunktionen für den empfangsseitigen Verkehr wurde bereits in Abschnitt 5.6.1 detailiert dargestellt. In Tabelle 5.10 werden die Ergebnisse zusammengefasst. Aufgrund der Kriterien kann die Lognormalverteilung als Verteilung mit der

	aggregierte Zwischenankunftszeiten				
	Parameter		d_{max}	r	RMS
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	8,3815	86,8754	0,0643	0,99866	0,0211
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	4,4014	0,1407	0,0126	0,99995	0,00586
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	9,7501	76,6011	0,0251	0,99641	0,00992
	summierte Paketlängen				
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	5,0040	48.357,08	0,0550	0,99725	0,0278
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	10,6755	0,2260	0,0135	0,99985	0,00643
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	6,4257	39.615,90	0,0418	0,9935	0,0210

Tabelle 5.10: geschätzte Parameter und Vergleich der Anpassung, Hochlast (empfangsseitg)

besten Anpassung angenommen werden. Das gilt sowohl für die Paketlängen als auch für die Zwischenankunftszeiten. Diese Aussage wird durch die zugehörigen „Normal Probability Plots“ und P-P-Plots in Abschnitt 5.6.1.2 untermauert. In Abbildung 5.18 werden die Dichtefunktionen der empirischen und theoretischen Verteilungen der aggregierten Zwischenankunftszeiten dargestellt. Zum Vergleich wird ebenfalls die Dichtefunktion der Poissonverteilung gezeigt. Zur Wahrung der Übersichtlichkeit wurde die Abszissenachse

logarithmisch geteilt. Die Abbildung zeigt, daß die Annahme der Lognormalverteilung als bestes Modell begründet ist, was mit den Ergebnissen in (LWDW97) übereinstimmt. Hier wurde ebenfalls der aggregierte Ankunftsprozess mittels Lognormalverteilung beschrieben.

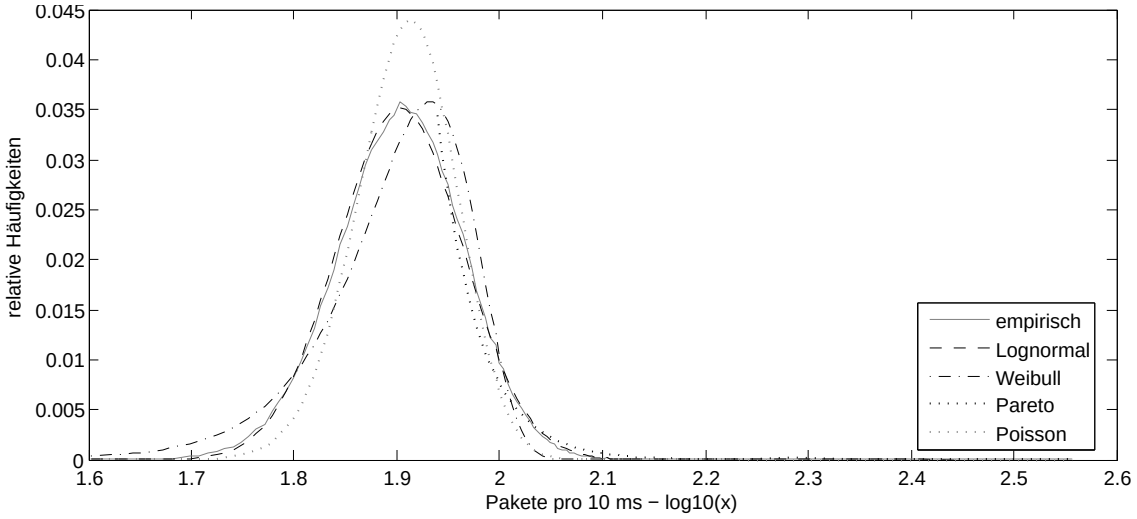


Abbildung 5.18: Vergleich der Verteilungen (agg. Zwischenankunftszeiten, empfangen, 18 bis 19 Uhr)

5.6.2.2 Ausgehender Verkehr

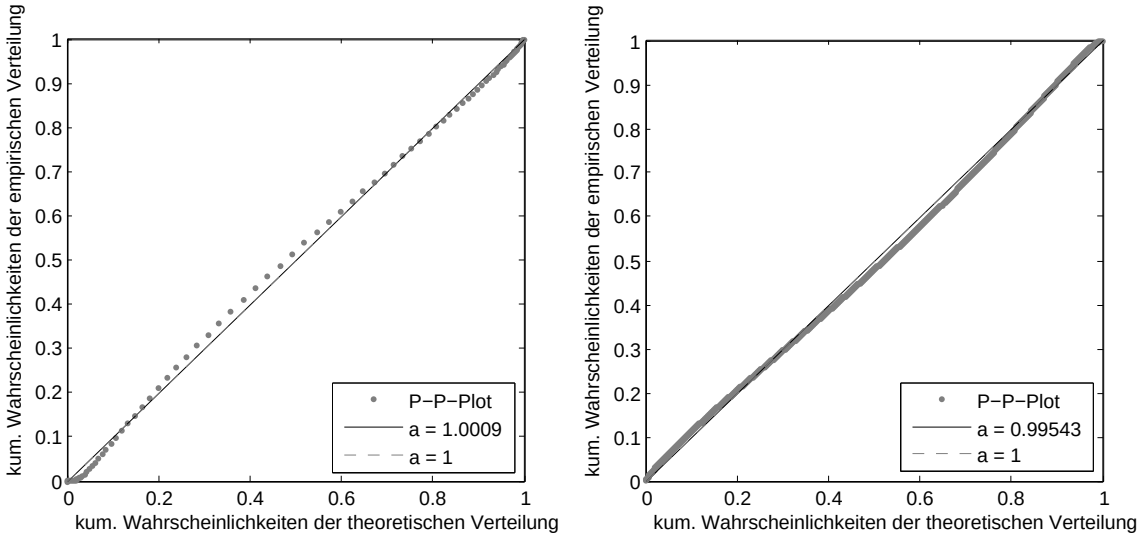
Für den ausgehenden Verkehr wurden alle zur Verfügung stehenden Schätzer verwendet. In diesem Abschnitt werden jedoch nur die geschätzten Parameter mit der jeweils größten Anpassung in Tabelle 5.11 zusammengefasst. In allen Fällen ermöglichten die Verteilungsparameter, die mittels Momentenschätzer bestimmt wurden, die größte Anpassung an die empirische Verteilungsfunktion. Der Tabelle ist zu entnehmen, daß die summierten Paket-

	aggregierte Zwischenankunftszeiten				
	Parameter		d_{max}	r	RMS
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	5,1143	78,3147	0,0157	0,99937	0,0484
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	4,2551	0,2210	0,0239	0,99973	0,0106
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	6,9570	65,5022	0,0355	0,99551	0,0137
	summierte Paketlängen				
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	3,9407	27.452,56	0,0238	0,99955	0,0115
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	10,08220	0,2789	0,0426	0,99852	0,0213
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	5,6926	22.038,37	0,0627	0,98819	0,0318

Tabelle 5.11: geschätzte Parameter und Vergleich der Anpassung, Hochlast (sendeseitg)

längen am besten mit der Weibullverteilung beschrieben werden (vgl. Abbildung 5.19(b)). Für die aggregierten Zwischenankunftszeiten ist die Lognormalverteilung das geeignete Mittel zur Beschreibung der empirischen Verteilungsfunktion. Verglichen mit den Werten

in Abschnitt 5.6.1.2 fallen die Werte für den Root Mean Square Error und den Korrelationskoeffizienten weniger eindeutig aus. Auch der P-P-Plot in Abbildung 5.19(a) zeigt diese Diskrepanz in Form einer Krümmung, die für die Weibull- und Paretoverteilung noch stärker ausfällt. Abschließend zu diesem Abschnitt werden in Abbildung 5.20 die theoretischen Verteilungsfunktionen mit der empirschen Verteilungsfunktion vergleichend dargestellt.



(a) ***agg. Zwischenankunftszeiten (Lognormal)*** **(b)** ***summierte Paketlängen (Weibull)***

Abbildung 5.19: Darstellung der P-P-Plots (ausgehender Verkehr, 18 bis 19 Uhr)

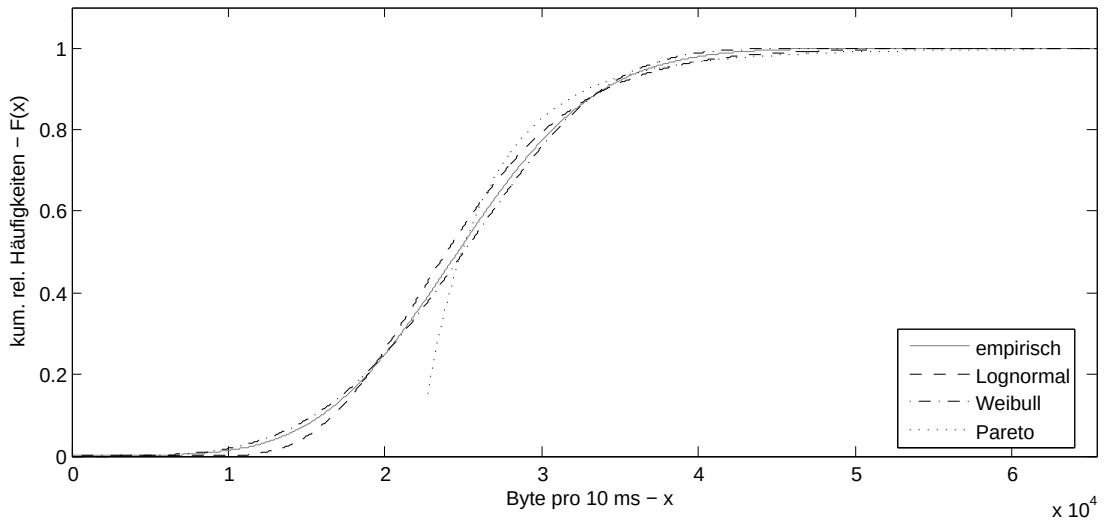


Abbildung 5.20: Vergleich der Dichtefunktionen (gesendete Paketlängen, 18 bis 19 Uhr)

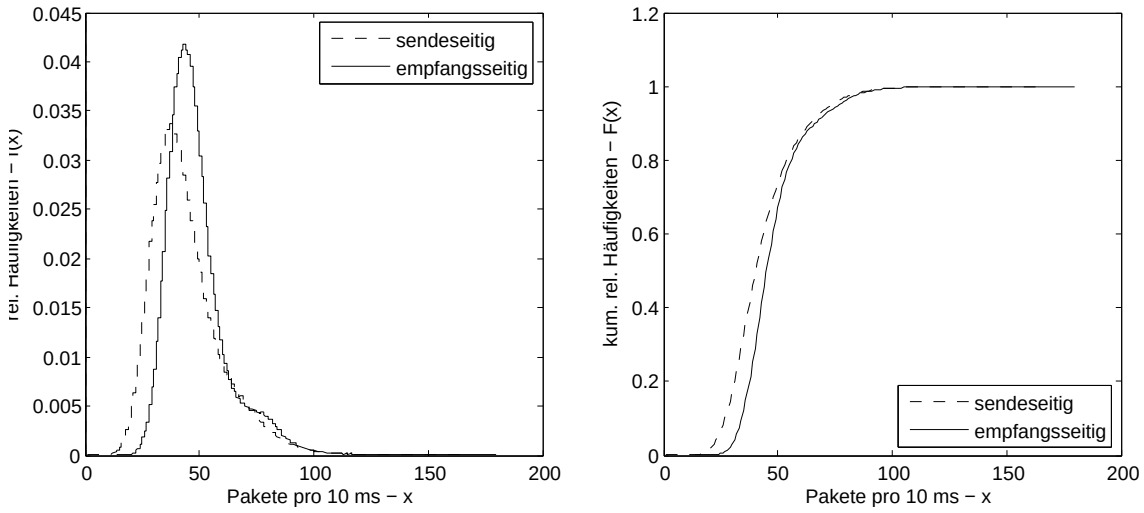
5.6.3 Niedriglastsituation

Während des betrachteten Bereichs der niedrigen Ausnutzung (zwischen 07 und 08 Uhr) wurden in 15.756.955 Paketen $4{,}82966 \cdot 10^9$ Byte gesendet und in 17.420.902 Paketen $7{,}99248 \cdot 10^9$ Byte empfangen. Gemessen am empfangen Datenvolumen zu Zeiten hoher Auslastung sind hier nur ca. 50% des Datenvolumens zu verzeichnen. Dementsprechend fallen auch die arithmetischen Mittelwerte und die Medianwerte der Zwischenankunftszeiten relativ hoch aus, wie in Tabelle 5.12 zu erkennen. Die minimal auftretenden Paket-

	Ankunftszeiten [μs]		Paketlängen [Byte]	
	empfangen	gesendet	empfangen	gesendet
arithmetisches Mittel	206,648	228,470	458,787	306,510
Median	137,731	95,889	60	29
Standardabweichung	227,221	340,517	569,172	498,977
Minimalwert	0,656	0,656	8	8
Maximalwert	4.263,67	17.916,81	1.480	1.480

Tabelle 5.12: Eigenschaften der Zwischenankunftszeiten und Paketlängen (Niedriglast)

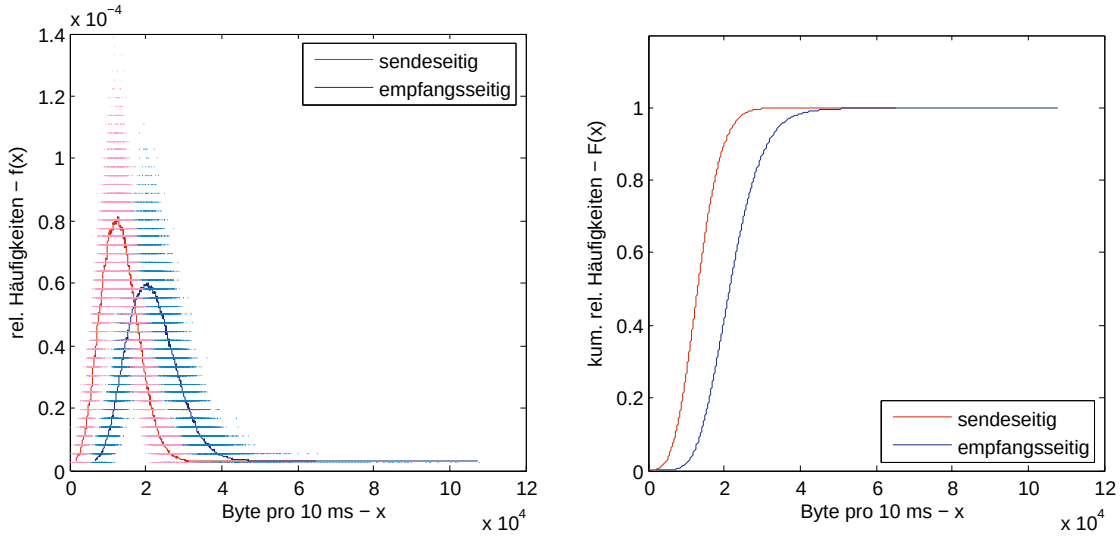
längen konnten mittels geeigneter Anfragen zu einem Großteil ICMP (Internet Control Message Protocol) zugeordnet werden. In Abbildung 5.21 werden die Wahrscheinlichkeitsfunktionen und Wahrscheinlichkeitsdichtefunktionen der aggregierten Zwischenankunftszeiten dargestellt. Die Dichte- und Verteilungsfunktionen der summierten Paketlängen



(a) *Verteilungsdichten* **(b) *Verteilungsfunktionen***

Abbildung 5.21: empirische Verteilungen der agg. Zwischenankunfszeiten (7 bis 8 Uhr)

sind in Abbildung 5.22 zu finden. Bereits bei der Betrachtung der Hochlastsituation waren die Dichtefunktionen der Paketlängen nicht klar definiert. Auch hier mußte mittels gleitendem Mittelwert (engl.:*moving average*) eine Glättung erfolgen.



(a) ***Verteilungsdichten*** **(b)** ***Verteilungsfunktionen***

Abbildung 5.22: empirische Verteilungen der summierten Paketlängen (7 bis 8 Uhr)

5.6.3.1 Eingehender Verkehr

Für die in diesem Abschnitt betrachteten Daten wurden wiederum alle zur Verfügung stehenden Schätzer der jeweiligen Verteilungen angewandt. Aus Gründen der Übersichtlichkeit werden in Tabelle 5.13 die Werte der Parameter mit der höchsten Anpassung zusammengefasst. Die oberen 50% der Verteilungsfunktion der aggregierten Zwischenan-

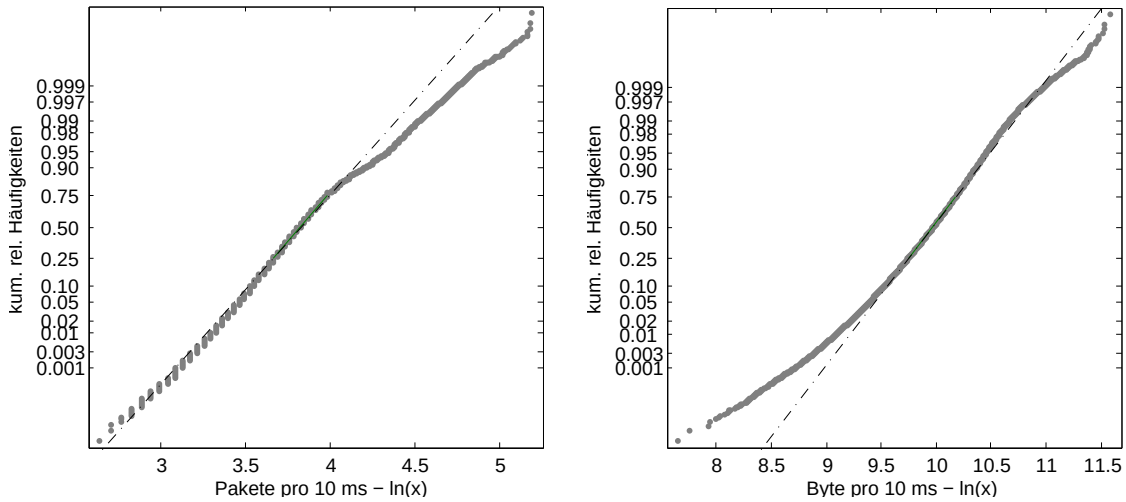
	aggregierte Zwischenankunftszeiten				
	Parameter		d_{max}	r	RMS
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	3,9526	52,9924	0,1235	0,99363	0,0402
Lognormal[2]($\hat{\mu}_L$,$\hat{\sigma}_L$)	3,8447	0,2571	0,0685	0,99845	0,0204
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	5,0981	41,1002	0,0616	0,99548	0,0107
	summierte Paketlängen				
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	3,3981	24.712,91	0,0461	0,99809	0,0238
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	9,9577	0,3169	0,0203	0,99967	0,0096
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	4,8650	18.996,10	0,0481	0,99140	0,0234

Tabelle 5.13: geschätzte Parameter und Vergleich der Anpassung, Niedriglast (empfangsseitg)

kunftszeiten werden durch die Paretoverteilung gut beschrieben. Über den Gesamtbereich gesehen beschreibt die Lognormalverteilung die Zwischenankunftszeiten und Paketlängen am besten. Während der Analyse trat hier der einzige Fall auf, bei dem der Maximum Likelihood Schätzer einen besseren Wert lieferte. Die sonstigen angegebenen Parameter wurden mit dem Momentenschätzer ermittelt.
In Abbildung 5.23 werden die „Normal Probability Plots“ dargestellt. Während für die Paketlängen eine Lognormalverteilung wahrscheinlich ist, tritt für die Zwischenankunfts-

2 mit MLE geschätzt

zeiten ein Versatz auf. Dieser Versatz ist in Abbildung 5.21(a) deutlich als „Buckel“ zu erkennen. Die hier beschriebene Abweichung erschwerte die Anpassung an die empirsche Verteilung, wie in Abbildung 5.24 gezeigt wird. Hier zeigt sich der MLE als robuster Schätzer, während der oben beschriebene „Buckel“ die statischen Momente verzerrt und somit der Momentenschätzer keine genauen Werte liefert.



(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung 5.23: Normal Probability Plot, 7 bis 8 Uhr, eingehender Verkehr

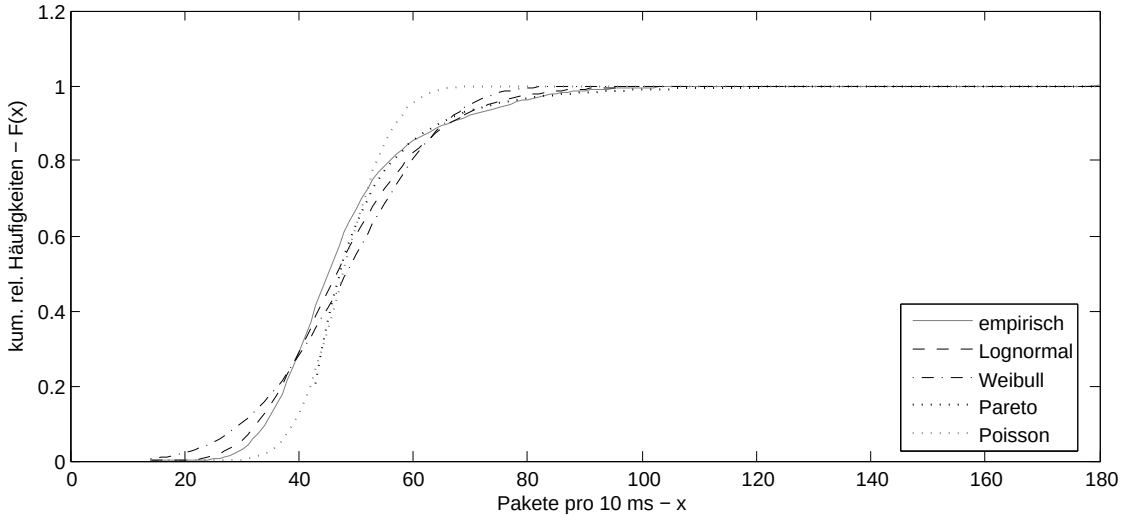


Abbildung 5.24: Vergleich der Verteilungen (agg. Zwischenankunftszeiten, empfangen, 7 bis 8 Uhr)

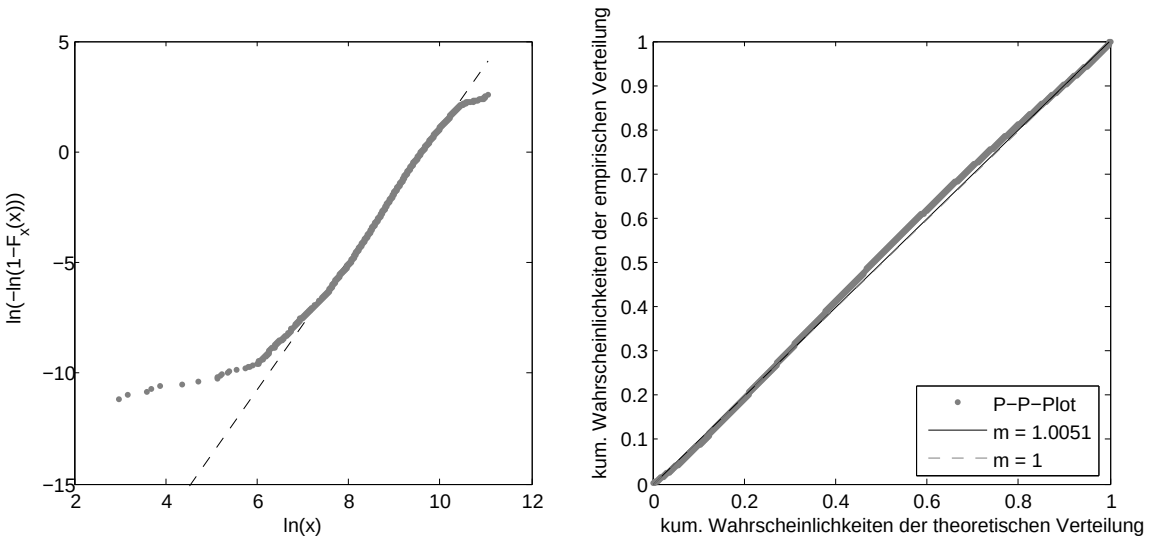
5.6.3.2 Ausgehender Verkehr

Für den ausgehenden Verkehr der Niedriglastsituation sind in Tabelle 5.14 die geschätzten Parameter zusammengefasst. Die Schätzungen der hier angegebenen Parameter erfolgten jeweils mittels Momentenschätzer, da dieser in allen Fällen den höchsten Grad der Anpassung ermöglichte. Wie bereits beim ausgehenden Verkehr unter hoher Auslastung,

	aggregierte Zwischenankunftszeiten				
	Parameter		d_{max}	r	RMS
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	3,1426	49,1687	0,1015	0,99609	0,0344
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	3,7210	0,3404	0,0331	0,99971	0,0100
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	4,4434	36,4088	0,0569	0,99177	0,0125
	summierte Paketlängen				
Weibull ($\hat{\alpha}_w$,$\hat{\beta}$)	2,8898	15.047,77	0,0199	0,99968	0,0098
Lognormal ($\hat{\mu}_L$,$\hat{\sigma}_L$)	9,4381	0,3635	0,0430	0,99842	0,0224
Pareto ($\hat{\alpha}_p$,$\hat{t}_0$)	4,5294	11.439,71	0,0670	0,98525	0,0332

Tabelle 5.14: geschätzte Parameter und Vergleich der Anpassung, Niedriglast (sendeseitg)

werden auch hier die summierten Paketlängen durch die Weibullverteilung sehr genau beschrieben. Dies war während der Analyse bereits aus der Darstellung für den LSE zu erkennen. Der zugehörige P-P-Plot bestätigte dies. Beide Darstellungen sind in Abbildung 5.25 aufgeführt. Von den hier betrachteten Verteilungen, wird der aggregierte Ankunfts-



(a) *Least Squares Estimator* **(b) *P-P-Plot***

Abbildung 5.25: LSE und P-P-Plot der summierte Paketlängen (Weibull)

prozess durch die Lognormalverteilung am besten beschrieben. Das entspricht ebenfalls dem Ergebnis aus Abschnitt 5.6.2.2. In Abbildung 5.26 werden die verschiedenen Dichtefunktionen vergleichend dargestellt. Zum Vergleich ist ebenfalls die Dichtefunktion der aus den Daten geschätzten Poissonverteilung abgebildet. Es ist zu erkennen, daß die Pois-

sonverteilung hier zu nicht brauchbaren Ergebnissen führt. Die Lognormalverteilung wird als Verteilung mit der höchsten Anpassung bestätigt.

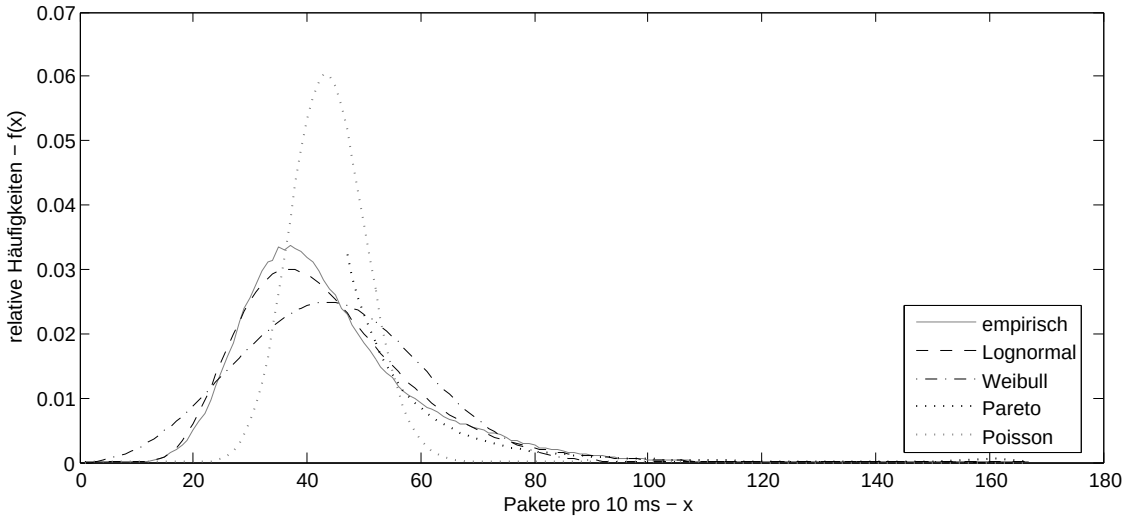


Abbildung 5.26: Vergleich der Dichtefunktionen (agg. Zwischenankunftszeiten, gesendet, 7 bis 8 Uhr)

5.7 Zusammenfassung der Ergebnisse der Analyse

In diesem Abschnitt werden die während der Analyse gewonnenen Ergebnisse zusammengefasst. Die Analysen wurden mit dem Programm MATLAB in der Version 7 (R14) durchgeführt.
Bereits bei der Betrachtung der analysierten Datensätze in Abschnitt 5.3 konnten aus den Abbildungen nicht-stationäre Einflüsse ermittelt werden. Dies konnte in Abschnitt 5.4 bestätigt werden, indem die Mittelwerte und Standardabweichungen in zwei Zeitbereichen gegenübergestellt wurden. Schlußfolglich wurden zwei Belastungssituationen untersucht.
Während des Zeitraums hoher Netzauslastung wurden mehr als 29 Millionen Pakete in einer Stunde empfangen. Die Analyse der reinen Zwischenankunftszeiten hätte die Grenzen der zur Verfügung stehenden Rechnerkapazitäten bei weitem überschritten. So war ein aggregierte Ankunftsprozess mittels geeigneter Verteilungsfunktionen zu beschreiben. Das führte dazu, daß die Exponentialverteilung aus der Analyse ausgeschlossen wurde. In einigen Abbildung wird jedoch die Poissonverteilung vergleichend hinzugezogen.
Bei der Bestimmung des Hurst-Parameters konnte gezeigt werden, daß der betrachtete Ankunftsprozess langzeitabhängig ist. Die Mittelwerte der Hurst-Parameter lagen zwischen $0,6726 \leq H \leq 0,7590$ und wichen signifikant von 0,5 ab. Dies war eine Vorraussetzung, um Langzeitabhängigkeit nachzuweisen (vgl. Abschnitt 1.2.3).
Bei der Schätzung der Verteilungsparameter bestätigte sich die Annahme der unzurei-

chenden Genauigkeit des Least Squares Estimators. Dessen Möglichkeit zur Erstellung von Prognosen konnte in Anschnitt 5.6.1 veranschaulicht werden. Für die eigentliche Bestimmung der Verteilungsparameter waren jedoch nur der Maximum Likelihood Estimator und die Methode der Momenten von Relevanz. Erwartet wurde, daß der MLE genauere Schätzungen liefern wird, als der Momentenschätzer. Dieser Fall ist jedoch bei 24 Anwendungen nur einmal aufgetreten. Die 24 ($2 \cdot 2 \cdot 3 \cdot 2$) Anwendungen ergeben sich aus:

- 2 Belastungssituationen
- 2 Übertragungsrichtungen
- 3 überprüfte Verteilungen
- Betrachtung der aggregierten Zwischenankunftszeiten und summierten Paketlängen

Hier zeigte der MLE seine robusten Eigenschaften, gegenüber Abweichungen (vgl. Abschnitt 5.6.3.1). In den sonstigen Fällen lieferte der Momentenschätzer Werte für die Parameter der theoretischen Verteilungsfunktion, die die genauste Anpassung an die empirische Verteilung ermöglichten. Wie bereits in Abschnitt 2.3 erwähnt, steigt die Genauigkeit des Momentenschätzers mit der Anzahl der Meßwerte.
Obwohl die gewählten Kriterien nur die Differenzen der theoretischen und empirischen Verteilungen bewerten (vgl. Kap. 3), konnten die Ergebnisse der Gütekriterien mit Hilfe der zugehörigen Darstellungen bestätigt werden. Kritisch zu bewerten ist die maximale Differenz d_{max}. Dieser Wert alleine ist ungeeignet, um Aussagen über den Grad der Anpassung zu erlauben.
Für die aggregierten Zwischenankunftszeiten der betrachteten Meßreihen konnte die Lognormalverteilung als Verteilung mit der höchsten Anpassung identifiziert werden. Dies entspricht den Ergebnissen in (LWDW97). Die Ergebnisse der Lognormalverteilung für den Datensatz des eingehenden Verkehrs zwischen 18 und 19 Uhr (Abschnitt 5.6.1) sind als Ausnahmen zu sehen. Ein solcher Grad der Anpassung konnte in den übrigen Datensätzen nicht erreicht werden.
Für die summierten Paketlängen ist das Ergebnis weniger eindeutig. Während die Paketlängen des eingehenden Verkehrs jeweils mit der Lognormalverteilung am besten beschrieben wurden, konnte beim ausgehenden Verkehr mit Hilfe der Weibullverteilung eine höhere Anpassung erzielt werden.

Kapitel 6

Zusammenfassung und Ausblick

Mit wachsenden Anforderungen an die Qualität von Netzwerken, gewinnen exakte Beschreibungen immer mehr an Bedeutung. In diesem Zusammenhang werden Modelle benötigt, die das bestehende Datenverkehrsaufkommen in dessen Charakteristika wiedergeben. Dies ermöglicht Rückschlüsse für die Realisierung von neuen Netzwerken. Um Zwischenankunftszeiten und Paketlängen effizient beschreiben zu können, werden Verteilungen benötigt, die diese Prozesse wiederspiegeln.
In Arbeiten wie (Fel01), (LTWW94) oder (LWDW97) wurde gezeigt, daß Ankunftsprozesse nicht weiter mittels Poisson- bzw. Exponentialverteilung beschrieben werden können, aufgrund der nachgewiesenen Langzeitanhängigkeit. Dies impliziert, daß die Annahme der Gedächtnislosigkeit nicht weiter gegeben ist. In dieser Arbeit wurde ebenfalls dieser Ansatz verfolgt. Hierzu wurden die Pareto-, Weibull- und Lognormalverteilung als „heavy-tailed"-Verteilungen untersucht. Vergleichend dazu wurde auch die Exponentialverteilung ausführlich bechrieben.
Bereits in Kapitel 1 wurden die genannten Verteilungen detailliert dargestellt. In Kapitel 2 wurden Methoden zur Bestimmung der Verteilungsparameter aufgezeigt. Die betrachteten Schätzverfahren sind der Least Squares Estimator, der Maximum Likelihood Estimator und die Methode der Momente.
Um die Anpassung zweier Verteilungen vergleichen zu können, wurden in Kapitel 3 Gütekriterien vorgestellt. Zu untersuchen waren dabei klassische Anpassungstests, wie der Kolmogorov-Smirnov-Anpassungstest und der χ^2-Anpassungstest. Diese geben jedoch kei-

ne Auskunft darüber wie gut eine empirische Verteilung mit einer theoretischen Verteilung übereinstimmt. Möglich ist nur die Aussage, ob eine empirische Verteilung durch eine gewählte theoretische Verteilung beschrieben wird. Hier ist die Gefahr von mehrdeutigen Ergebnissen zu sehen, wenn eine empirische Verteilung durch mehrere theoretische Verteilungen beschrieben wird und jeweils die Nullhypothese H_0 angenommen werden muß. Folglich wurden in Abschnitt 3.6 alternative Kriterien vorgestellt, mit denen die Anpassung zwischen theoretischer und empirischer Verteilung auch quantitativ bestimmt werden kann. Hierzu zählen der Korrelationskoeffizient r und der Root Mean Square Error (kurz. RMS). Als graphische Methode zu Darstellung des Zusammenhangs zweier Verteilungen wurde der P-P-Plot verwendet.
Um die Annahme der Langzeitabhängigkeit zu bestätigen, wurden in Kapitel 4 Methoden zur Bestimmung des Hurst-Parameters vorgestellt. Als Verfahren sind zu nennen: der Variance-Time-Plot, die R/S-Statistik und das Periodogramm. Anwendung fanden die oben genannten Methoden und Verteilungen in Kapitel 5. Betrachtet wurde das Datenverkehrsaufkommen der Universität Rostock über einen Zeitraum von ca. 24 Stunden. Dabei war zunächst auf die betrachteten Meßreihen und das Meßsystem einzugehen. Über den Gesamtbereich der Messung konnte von keiner Stationarität ausgegangen werden, da das Datenverkehrsaufkommen tageszeitlichen Schwankungen unterliegt (vgl. Abschnitt 5.4). Die Annahme der Stationarität gilt nur für einige Minuten bzw. eine Stunde.
Analysiert wurden die aggregierten Zwischenankunftszeiten und summierten Paketlängen für zwei Belastungssituationen von jeweils einer Stunde Dauer. Da es sich um aggregierte Ankunftsprozesse handelte, war die Exponentialverteilung von der Analyse auszuschließen. Aufgrund des Verlaufs der Verteilungsfunktion der Paretoverteilung (vgl. Abb. 1.5) war davon auszugehen, daß nur eine Anpassung in den oberen 50% der Verteilung zu erreichen ist. In nachfolgenden Arbeiten könnte die Anwendung der hier beschriebenen Verfahren und Verteilungen auf reine Zwischenankunftszeiten interessant sein.
Die Analyse des aggregierten Netzwerkverkehrs ergab eindeutige Hinweise auf Langzeitabhängigkeit (vgl. Abschnitt 5.5). Dabei traten jedoch große Schwankungen zwischen den geschätzten Hurst-Parametern auf. An dieser Stelle könnten wiederum weiterführende Arbeiten ansetzen, die die Nutzung weiterer Verfahren beinhaltet. Weiterhin ergab die Analyse, daß die aggregierten Zwischenankunfszeiten in allen Meßreihen lognormalverteilt sind mit teilweise sehr hohen Anpassungen (Bsp.: Abschnitt 5.6.1.2). Dieses Ergebnis entspricht den Resultaten in (LWDW97). Für die summierten Paketlängen gilt beim eingehenden Verkehr ebenfalls die Annahme der Lognormalveteilung. Für den ausgehenden Verkehr wurde die Weibullverteilung als Verteilung mit der höchsten Anpassung ermittelt. Eine ausführlichere Darstellung der Analyseergebnisse ist in Abschnitt 5.7 zu finden. In weiteren Arbeiten könnte ebenfalls der Betrachtungszeitraum vergrößert oder die Betrachtungen auf die übertragenen Dateigrößen erweitert werden.

Literaturverzeichnis

[A99] Henrik Abrahamsson
Traffic measurment and analysis
URL: http://www.sics.se/∼henrik/t9905.pdf (Datum: 14.01.05)

[AS86] Ralph B. D'Agostino, Michael A. Stephens
Goodness-of-fit Techniques Schriftenreihe: Statistics - Textbooks and Monographs
Vol. 68, Dekker, New York, ISBN 0-8247-7487-6, 1986

[Bar98] Hans-Jochen Bartsch
Taschenbuch mathematischer Formeln, 18. verbesserte Auflage
Carl Hanser Verlag, ISBN 3-446-19396-0, München 1998

[BD91] Peter J. Brockwell, Richard A. Davis
Time Series: Theory and Methods, Second Edition
Springer Verlag, ISBN 0-387-97429-6, New York 1991

[BG99] Josef Bleymüller, Günther Gehlert
Statistische Formeln, Tabellen und Programme, 9. überarbeitete Auflage
Verlag Franz Vahlen, ISBN 3-8006-2463-X, München 1999

[Bor01] Michael S. Borella
On Estimating Long Range Dependence of Network Delay
International Journal of Chaos Theory and Applications, Vol. 6, Nummer 4, 2001

[Bri01] David R. Brillinger
Time Series: Data Analysis and Theory
Society for Industrial and Applied Mathematics, ISBN 0-89871-501-6, 2001

[Bru05] Udo Brunswig
Diplomarbeit zum Thema: *Analyse und Visualisierung des Kommunikationsaufkommens in paketorientierten Netzen*, Universität Rostock, Januar 2005

[BSMM99] I. N. Bronstein, K. A. Semendjajew, G. Musiol, H. Mühlig
Taschenbuch der Mathematik
4. überarbeitete und erweiterte Auflage der Neubearbeitung, Verlag Harri Deutsch, ISBN 3-8171-2004-4, 1999

[CB95] Mark Crovella, Azer Bestavros
Explaining World Wide Web Traffic Self-Similarity
1995; URL http://citeseer.ist.psu.edu/crovella95explaining.html (Datum 19.11.04)

[Den96] Shuang Deng *Empirical Model of WWW Document Arrivals at Access Link*
IEEE International Conference on Communication, June 1996,
URL: http://citeseer.ist.psu.edu/deng96empirical.html (Datum 25.12.04)

[DOT02] Paul Doukhan, George Oppenheim, Murad S. Taqqu
Theory and Applications of Long-Range Dependence
Birkhäuser Verlag, ISBN 0-8176-4168-8, Boston 2002

[Dow01] Allen Downey
Evidence for long-tailed distributions in the Internet
ACM SIGCOMM Internet Measurement Workshop, November 2001, URL http://citeseer.ist.psu.edu/downey01evidence.html (Datum 19.11.04)

[Dut03] R. Dutter
begleitendes Material der Vorlesung: *Statistik und Wahrscheinlichkeitsrechnung*
Technische Universität Wien, Wintersemester 2003
URL: http://www.statistik.tuwien.ac.at/public/dutt/vorles/mb/mb_wi_vt.html (Datum 15.01.05)

[Fel01] Anja Feldmann
Characteristics of TCP Connection Arrivals
Self-Similar Network Traffic and Performance Evaluation, John Wiley & Sons, Inc., 2001 ISBN 0-471-31974-0

[Gess93] Jürgen R. Geßler
Statistische Grafik
Birkhäuser Verlag, ISBN 3-7643-2874-6, Basel 1993

[GK04] Christian Groth, Jens Kosubek
Studienarbeit zum Thema: *Modellierung und Simulation von Ethernet-Netzwerkverkehr*, Universität Rostock 2004

[Hab02] Marco Haberland
Klausurwissen Statistik, Technische Universität Hamburg
URL: http://www.tu-hamburg.de/ sbmh0592/studium/klausurwissen_statistik.pdf (Datum: 09.12.04)

[HEK93] Joachim Hartung, Bärbel Elpelt, Karl-Heinz Klösener
Statistik: Lehr- und Handbuch der angewandten Statistik, 9. durchgesehene Auflage
Oldenbourg Verlag, ISBN 3-486-22658-4, München Wien 1993.

[JKB70a] Norman Lloyd Johnson, Samuel Kotz, Narayanaswamy Balakrishnan
Continuous Univariate Distributions - 1

Wiley series in probability and mathematical statistics : Applied probability and statistics; Distributions in Statistics, John Wiley & Sons, Inc., 1970 ISBN 0-471-44626-2

[JKB70b] Norman Lloyd Johnson, Samuel Kotz, Narayanaswamy Balakrishnan
Continuous Univariate Distributions - 2
Wiley series in probability and mathematical statistics : Applied probability and statistics; Distributions in Statistics, John Wiley & Sons, Inc., 1970 ISBN 0-471-44627-0

[JKB70c] Norman LLoyd Johnson, Samuel Kotz, Narayanaswamy Balakrishnan
Discrete Distributions
Wiley series in probability and mathematical statistics : Applied probability and statistics; Distributions in Statistics, John Wiley & Sons, Inc., 1970 ISBN 0-471-44360-3

[Kes05] Thomas Kessler
(noch unveröffentlichte) Dissertation: *Ein Beitrag zur Messung und Charakterisierung multimedialer Datenströme in IP Zugangsnetzen*
Universität Rostock

[KFR02] Thomas Karagiannis, Michalis Faloutsos, Rudolf H. Riedl
Long-Range Dependence: Now you see it, now you don't!
Rice University 2002, URL: http://cmc.rice.edu/docs/docs/Kar2002Nov5Long-Rang.pdf (Datum: 15.02.05)

[KMV03] T. Kessler, H.-D. Melzer, T. Vergin
Analyse des Verkehrsaufkommens in hochbitratigen Zugangsnetzen
In: 4. IuK-Tage Mecklenburg-Vorpommern, Rostock, 18.-20. Juni 2003

[LWDW97] Mattew T. Lucas, Dallas E. Wrege, Bert J. Dempsey, Alfred C. Weaver
Statistical Characterization of Wide-Area IP Traffic
Proceedings of Sixth International Conference on Computer Communications and Networks, Seiten 442-447, 1997

[LTWW93] Will E. Leland, Murad S. Taqqu, Walter Willinger, Daniel V. Wilson
On the Self-Similar Nature of Ethernet Traffic
Notices of the American Mathematical Society; ACM SIGCOMM Computer Communication Review; vol. 23 nr. 4; Oktober 1993

[LTWW94] Will E. Leland, Murad S. Taqqu, Walter Willinger, Daniel V. Wilson
On the Self-Similar Nature of Ethernet Traffic (Extended Version)
IEEE/ACM Transactions on Networking; Februar 1994

[MVW93] Murad S. Taqqu, Vadim Teverovsky, Walter Willinger
Estimators for Long-Range Dependence: An Empirical Study
Fractals, 3(4):785-788, 1995

[Nie04] Heiko Niedermayer
begleitendes Material der Vorlesung: *Grundlagen zu Kommunikationsnetzen und Warteschlangentheorie*, Universität Tübingen 2004,
URL: http://net.informatik.uni-tuebingen.de/teaching/kommnetze/pdf_ws04/GrundlagenUndWarteschlangen.pdf (Datum: 27.02.05)

[Pax94] Vern Paxson
Empirically derived analytic models of wide-area TCP connections
IEEE/ACM Transactions on Networking, 2(4):316-336, 1994

[Pax97] Vern Paxson
Fast, Approximate Generation of Fractional Gaussian Noise for Generating Self-Similar Network Traffic
Computer Communications Review, V. 27 N. 5, October 1997, pp. 5-18

[PF95] Vern Paxson, Sally Floyd
Wide area traffic: The failure of poisson modeling
IEEE/ACM Transactions on Networking, 3(1):226-244, 1995

[Ryu96] B. K. Ryu
Fractal Network Traffic: From Understanding to Implications
PhD thesis, Columbia University, 1996

[Ryu00] B. K. Ryu
A Tutorial on Fractal Traffic Generators in OPNET for Internet Simulation
OPNETWORK 2000.

[Ryu02] B. K. Ryu
Fractal Traffic Models for Internet Simulation - Präsentation 2002
URL: http://www.koren21.net/workshop/2002/pdf/524-9.pdf (Datum: 23.11.05)

[RL98] B. K. Ryu, S. Lowen
Point Process models for self-similar network traffic, with applications
Stochstaic Models, 14(3):731-761,1998

[Ros96] O. Rose
Estimation of the Hurst Parameter of Long-Range Dependent Time Series
Universität Würzburg 1996, URL: http://citeseer.ist.psu.edu/rose96estimation.html (Datum: 15.02.05)

[RS02] Horst Rinne, Katja Specht
Zeitreihen: Statistische Modellierung, Schätzung und Prognose
Verlag Vahlen, ISBN 3-8006-2877-5, München 2002

[RSC04] ReliaSoft Corporation
Maximum Likelihood Estimation
URL: http://www.weibull.com/LifeDataWeb/
maximum_likelihood_estimation_appendix.htm (Datum: 11.12.04)

[Sch01] Stefanie Christina Scheid
Diplomarbeit zum Thema: *Die verallgemeinerte Lognormalverteilung*
Universität Dortmund, November 2001,
URL: http://compdiag.molgen.mpg.de/docs/scheid_diplom.pdf (Datum 18.01.05)

[Sti01] Winfried Stier
Methoden der Zeitreihenanalyse
Springer-Verlag, ISBN 3-540-41700-1, Berlin 2001

[Tan00] Andrew S. Tanenbaum
Computernetzwerke
3. revidierte Auflage, Prentice Hall, ISBN 3-8273-7011-6, München 2000

[WH87] Neil A. Weiss, Matthew J. Hassett
Introductory Statistics, second edition,
Addison-Wesley Publishing Company, ISBN 0-201-09582-3, 1987

[WP98] Walter Willinger, Vern Paxson
Where Mathematics meets the Internet
Notices of the American Mathematical Society; vol. 45 nr. 8; Seiten 961-970; 1998

[WTSW97] Walter Willinger, Murad S. Taqqu, Robert Sherman, Daniel V. Wilson
Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level
IEEE/ACM Transactions on Networking, Vol. 5, No. 1, 1997, Seiten 71-86.

Anhang A

weiterführende Abbildungen

In Abschnitt 5.3 wurde die Darstellung der eingehenden Pakete pro Sekunde aus Gründen der Übersichtlichkeit begrenzt. Im eine Vollständigkeit zu gewährleisten, wird in Abbildung A.1 der Gesamtbereich dargestellt. Die hier gezeigt schwarze Linie entspricht bereits einer Glättung mit Hilfe eines gleitenden Mittelwerts.

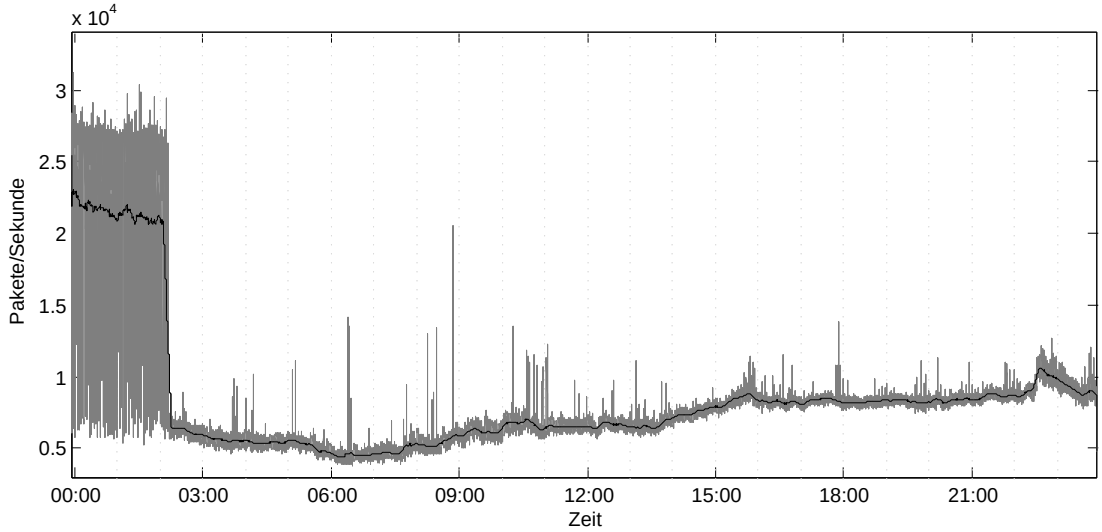
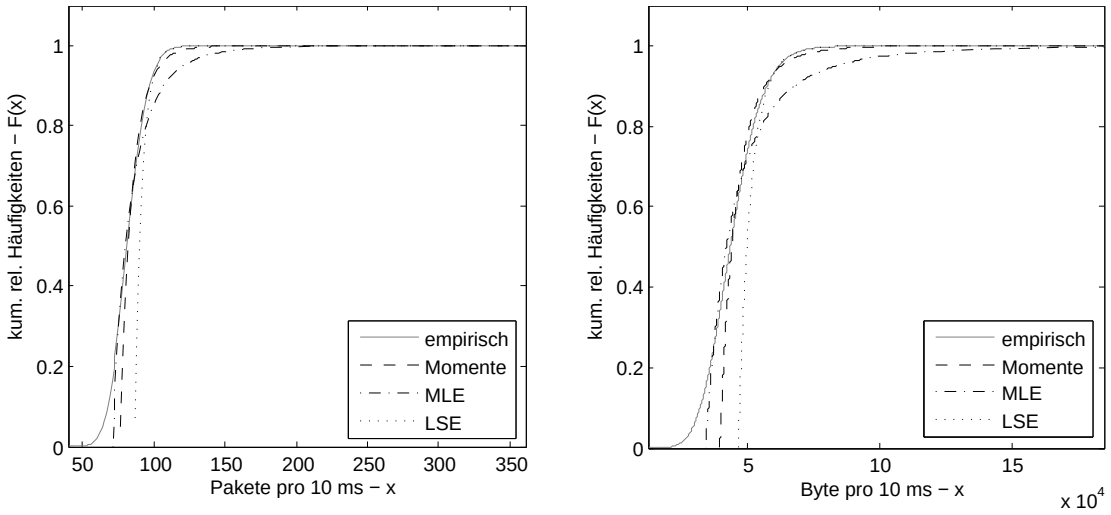


Abbildung A.1: Darstellung der eingehenden Pakete in Abhängigkeit von der Zeit (Gesamtbereich)

In Abschnitt 5.6.1.3 wurde die Schätzung der Paretoverteilungsparameter detailliert dargestellt. Der Vergleich der geschätzten Verteilungsparameter wurde auf die oberen 50% der Verteilung begrenzt. In Abbildung A.2 wird die Darstellung über den Gesamtbereich gezeigt.Es ist zu erkennen, daß der obere Bereich gut mit dem Momentenschätzer beschrieben werden kann. Der untere Teil der Verteilung wird jedoch vernachlässigt.



(a) *aggregierte Zwischenankunftszeiten* **(b) *summierte Paketlängen***

Abbildung A.2: Vergleich der Verteilungsparameter über den Gesamtbereich (Pareto)

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Rostock, den 5. April 2005

Jens Kosubek