Charakterisierung des Datenverkehrs von HTTP-Clienten

Dietrich Thie, Jens Flohrer
Professur Rechnernetze und verteilte Systeme,
Professur Modellierung und Simulation
Straße der Nationen 62, 09107 Chemnitz
E-mail: {dth,jfl}@informatik.tu-chemnitz.de

14. Januar 1999

Zusammenfassung

Der Entwurf von Mechanismen zur Verkehrsbeeinflussung in Datennetzen setzt statistisch gesicherte Modelle für die zu übertragenden Datenströme voraus. Der vorliegende Beitrag führt ein neues Modell für den Datenverkehr von HTTP-Clienten ein. Für die wichtigsten Modellparameter werden empirische Verteilungsgesetze und ihre Parameter bestimmt und diskutiert.

1 Einführung

Die ständige Steigerung des Internet-Datenaufkommens wird maßgeblich durch das starke Anwachsen der WWW-Nutzung verursacht. Dies macht eine Charakterisierung des HTTP-Datenverkehrs zur vorausschauenden Netzplanung erforderlich. Bisherige Untersuchungen befassen sich nur mit speziellen TCP-Diensten, wie FTP, Telnet, SMTP ([Pax94], [CBC95]) und eingeschränkt mit HTTP ([AW96], [CDJM91]). Diese Arbeit untersucht mit Hilfe eines neuen Modells den HTTP-Datenverkehr auf empirische Verteilungsfunktionen für die Analyse und Simulation.
Zur Analyse des HTTP-Datenverkehrs wurden die Log-Dateien des WWW-Servers der TU-Chemnitz im Zeitraum Januar bis Juni 1998 ausgewertet, da in dieser Zeit die Anforderungen an den Server dramatisch zugenommen haben. Die Untersuchungsergebnisse konnten unter Auswertung von Log-Dateien der TH-Zwickau und der TU-Dresden verifiziert werden. Zeitliche Szenarien waren bei diesen Untersuchungen nicht von Interesse, da sie von der jeweiligen Netzlastsituation, dem Caching und der Antwortzeit des Servers abhängig sind.

2 Modellbildung

In dieser Arbeit wurden Dateien, die der HTTP-Client vor dem Anzeigen einer HTML-Seite anfordert, als **direkte Dateien** bezeichnet, wie z.B. Bilder und Applets. Dateien, die der Nutzer durch einen Klick auf entsprechende Schaltflächen selbst anfordert, wurden als **indirekte Dateien** bezeichnet, unterschieden in **HTML-Dateien** und **others-Dateien**. Da die alleinige Auswertung der Log-Dateien keine Möglichkeit gab, z.B. Bilddateien als tatsächlich mittels HTML-Tag angeforderte direkte Dateien und nicht per Nutzeraktion angeforderte Datei zu erkennen, wurden folgende praktikable Konzessionen getroffen.

direkte Dateien	gif, jpg, jpeg, tif, pcx, xbm, class, java
HTML-Dateien	htm, html, shtml
others-Dateien	alle nicht genannten

Tabelle 1: Klassifizierung angeforderter Dateien

Den Modellmittelpunkt bildet eine HTML-Seite (Referer), auf die sich alle weiteren Betrachtungen beziehen. Mit einer bestimmten Wahrscheinlichkeit p_o wird der Nutzer des Web-Browsers einen eventuell vorhandenen Link in der Seite nutzen und eine indirekte Datei mit einer gewissen Größe S_o anfordern (FTP-likeVerfahren). Natürlich ist es mit einer weiteren Wahrscheinlichkeit p_h möglich, daß der Nutzer einem anderen Verweis folgt, um eine referenzierte HTML-Seite S_h anzufordern. Diese HTML-Seite veranlaßt den Web-Browser wiederum eine bestimmte Anzahl X_d von direkten Dateien der mitteleren Größe ES_d mitzuladen. Beide Nutzerentscheidungen p_o und p_h enden im Modell wieder in einer HTML-Seite. Dies bedeutet, daß eine angeforderte HTML-Seite erneut den Mittelpunkt des Modells bildet. Der Prozeß wird durch das Schließen des HTTP-Clienten beendet (p_e) bzw. aufgrund der Neueingabe eines weiteren URI (mit dem Ziel einer HTML-Seite) neu begonnen (p_s). Die beiden Wahrscheinlichkeiten p_s und p_e wurden in dieser Arbeit nicht betrachtet. Sie lassen sich jedoch aus Analysen der mittleren Bedienzeit gewinnen. Im weiteren wird angenommen: $p_o + p_h = 1$. Für Unterscheidungen der anfordernden Hosts wurde der jeweilige Bezeichner um einen zweiten Index ('l' = lokal oder 'g' = global) erweitert.

3 Ermittlung der Modellparameter

3.1 Entscheidungswahrscheinlichkeiten p_h und p_o

Zur Ermittlung der Modellparameter wurden nur Verbindungsanforderungen ausgewertet, die vom Server mit Status 2xx oder 3xx quittiert wurden. Da es bei

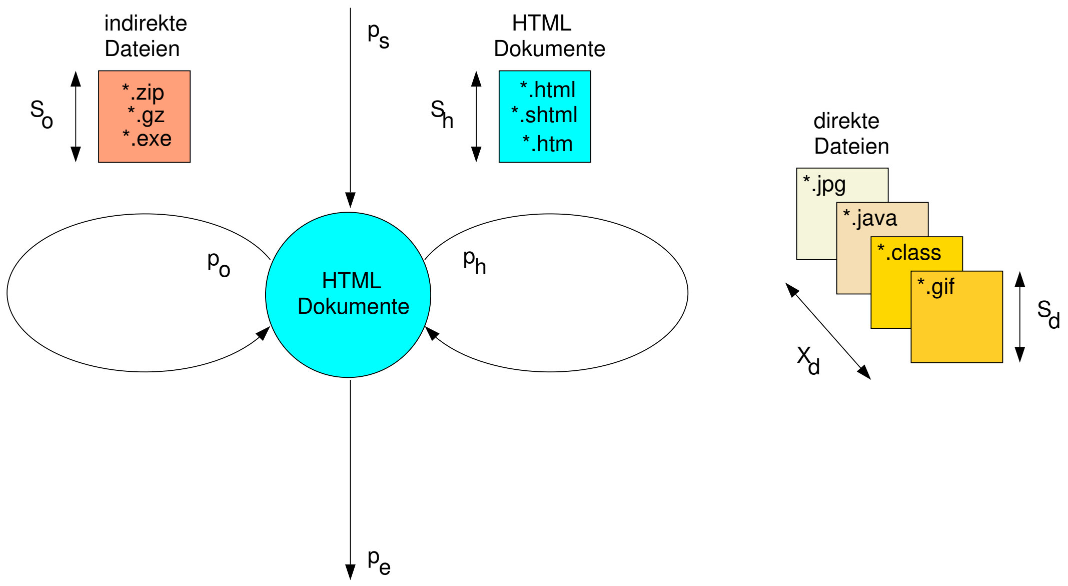

Abbildung 1: Modell des Datenverkehrs eines HTTP-Clienten

den untersuchten Wahrscheinlichkeiten starke Abweichungen in Bezug auf den rufenden Host gibt, wurde eine Unterscheidung in lokale Anforderungen (p_{hl}, p_{ol}) und globale Anforderungen (p_{hg}, p_{og}) eingeführt.

p_{hl}	p_{ol}	p_{hg}	p_{og}	p_h	p_o
96,29 %	3,71 %	85,26 %	14,75 %	89,93 %	10,07 %

Tabelle 2: Wahrscheinlichkeiten für die Nutzerentscheidung

Tabelle 2 zeigt deutlich die Dominanz der Nutzung von Verweisen zu weiteren HTML-Seiten gegenüber der Anforderung von others-Dateien, was ja als typisches Nutzerverhalten zu erwarten war. Die durchgeführte Unterscheidung zwischen lokalen und globalen Nutzern zeigt außerdem deutlich, daß sich das Anwenderverhalten in unterschiedlichen Nutzungsprofilen niederschlägt.

3.2 Verteilungsgesetz für die Anzahl direkter Dateien X_d

Zur Ermittlung des der Modellvariablen X_d zugrundeliegenden Verteilungsgesetzes wurde die Anzahl direkter Dateien von insgesamt ca. 6 Millionen Referer-Files ausgewertet.
Das Histogramm der Beobachtungswerte verdeutlicht zunächst einen relativ großen Variationskoeffizienten : Während einerseits ca. 90% der übertragenen Files weniger als 5 direkte Dateien beinhalteten, wurden andernseits auch extrem hohe Werte beobachtet. So wiesen z.Bsp. 0.012% der Referer mehr als 100 direkte Dateien auf, als Maximalwert wurden 458 direkte Dateien zu einem übertragenen

File registriert. Da gerade diese Ausreißer als mögliche Ursache für Staus und Übertragungsprobleme anzusehen sind, ist es im Kontext der behandelten Problemstellung sehr wichtig, daß das bei der Simulation benutzte Verteilungsgesetz auch diese Extreme korrekt abbildet.

Abbildung 2 zeigt die empirische Verteilungsfunktion $\hat{F}(x)$ in transformierter

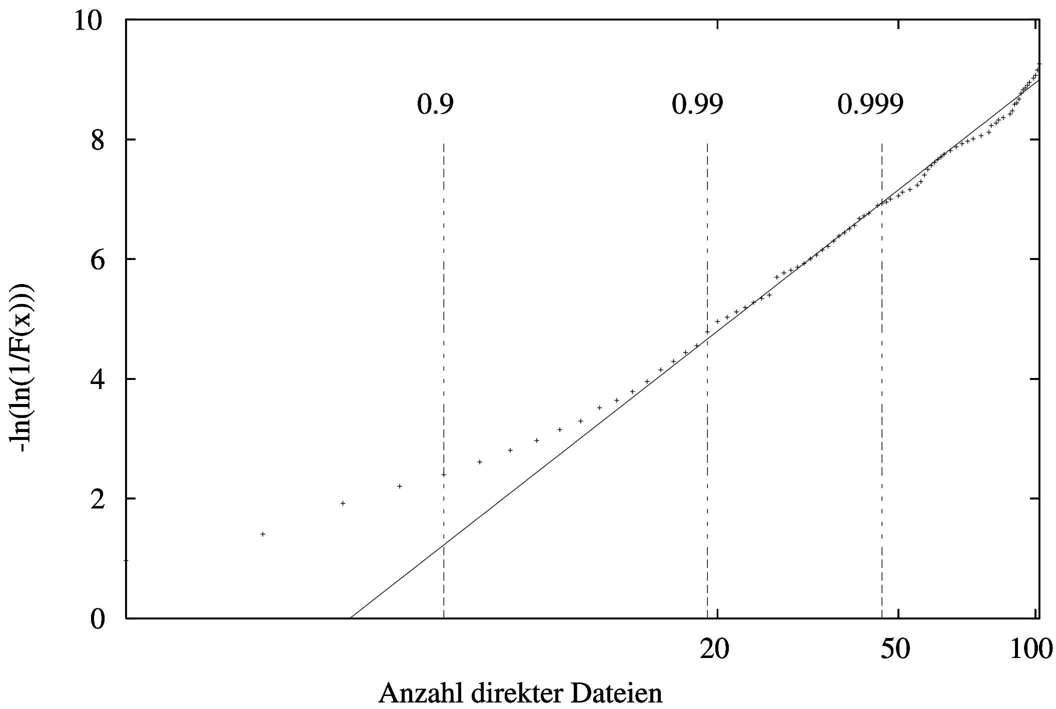

Abbildung 2: Transformierte empirische Verteilungsfunktion für X_d

Darstellung. Die senkrechten Linien markieren Levels der empirischen Verteilungsfunktion $\hat{F}(x)$, die an den betreffenden Stellen erreicht werden. Ab $x \geq 15$ verläuft $\hat{F}(x)$ in der gewählten Darstellung (Abszisse: Beobachtungswerte x im logarithmischen Maßstab; Ordinate: $1/\hat{F}(x)$, doppelt logarithmiert) annähernd linear; im Grenzverhalten entspricht das gesuchte Verteilungsgesetz somit einer Extremwertverteilung, dem Fisher-Tippett-Gesetz Typ II ([AS64] [Gum66]).

Die ersten 10 Punkte liegen alle oberhalb der Ausgleichsgeraden durch die restlichen Punkte: Je kleiner das Argument x , um so deutlicher der Abstand. Das bedeutet, daß vorgegebene Levels der Verteilungsfunktion im Vergleich zum Fisher-Tippett-Gesetz früher erreicht werden. Diese Erscheinung muß mittels einer Phasenverschiebung φ korrigiert werden. Die anzusetzende Verteilungsfunktion lautet somit

$$F(x) = \exp\left(-a(x+\varphi)^{-b}\right) \qquad (1)$$

Zur Schätzung der 3 Verteilungsparameter a, b und φ stehen zwei Möglichkeiten zur Auswahl:

- Zuerst Beschaffung einer Vorabschätzung $\hat{\varphi}$ für φ , dann Bestimmung von a und b aus den Parametern der Regressionsgeraden durch die Punktmenge ($ln(x_i + \hat{\varphi}), -lnln(1/\hat{F}(x_i))$).

- Direkte MKQ-Schätzung von a und b unter Elimination von φ . Anschließend Bestimmung von $\hat{\varphi}$ aus den Residuen zur inversen Verteilungsfunktion.

Variante 2 hat den Vorteil, daß der Gültigkeitsbereich des Fisher-Tippett-Gesetzes besser abgegrenzt werden kann. Die Elimination von φ gelingt dabei auf folgende Weise:
Zu jedem Argumentwert x aus dem Definitionsbereich der Verteilungsfunktion F ist über die Beziehung

$$F(u) = (F(x))^{1/\kappa} \tag{2}$$

ein „Partnerwert“ $u = u(x, \kappa)$ bezüglich einer passend vorgegebenen Konstanten $\kappa > 1$ definiert. Für die Differenz $u(x, \kappa) - x$ gilt dann unter Beachtung von (1) und (2) und mit $\beta = 1/b$

$$\begin{aligned} u - x &= \left(\left(\frac{a}{-lnF(u)} \right)^{\beta} - \varphi \right) - \left(\left(\frac{a}{-lnF(x)} \right)^{\beta} - \varphi \right) \\ &= a^{\beta} \left(\left(ln\frac{1}{F(u)} \right)^{-\beta} - \left(ln\frac{1}{F(x)} \right)^{-\beta} \right) \\ &= a^{\beta} \left(\left((-lnF(x))\frac{1}{\kappa} \right)^{-\beta} - (-lnF(x))^{-\beta} \right) \\ &= a^{\beta} \left(\left(\kappa^{\beta} - 1 \right) (-lnF(x))^{-\beta} \right) \end{aligned} \tag{3}$$

Die hergeleitete Gleichung kann zur Schätzung von a und b herangezogen werden. Zu fast allen Beobachtungswerten x_i, für die der Wert der empirischen Verteilungsfunktion $\hat{F}(x_i)$ bekannt ist, kann der Partnerwert $u_i = u(x_i)$ mittels Interpolation hinreichend genau bestimmt werden : u_i ist der interpolierte Argumentwert, bei dem die empirische Verteilungsfunktion das Niveau $(\hat{F}(x_i))^{1/\kappa}$ erreicht. Falls der Verteilungsansatz (1) korrekt ist, dann müßten die Punkte mit den Koordinaten $(-lnln(1/\hat{F}(x_i), ln(u_i - x_i))$ wegen (3) annähernd auf einer Geraden mit dem Anstieg β und dem Absolutglied $\beta ln(a) + ln(\kappa^{\beta} - 1)$ liegen.
Aus (3) läßt sich auch eine Abschätzung für die Phasenverschiebung φ herleiten. Da der Ausdruck $a^{\beta}/(-lny)^{\beta} - \varphi$ gerade die inverse Verteilungfunktion zu (1) repräsentiert, folgt aus (3) unmittelbar $u - x = (\kappa^{\beta} - 1)(x + \varphi)$ und daraus

$$x = \frac{1}{\kappa^{\beta} - 1}(u - x) - \varphi \tag{4}$$

Die Punkte $(u_i - x_i,\ x_i)$ müssen somit im Gültigkeitsbereich des Fisher-Tippett-Gesetzes auf einer Geraden liegen, deren Ordinatenschnittpunkt den Wert $-\varphi$ markiert. Abbildung 3 zeigt für $\kappa = 4$, daß die empirischen Daten die vermutete Eigenschaft tatsächlich nahezu im gesamten Wertebereich aufweisen[1].

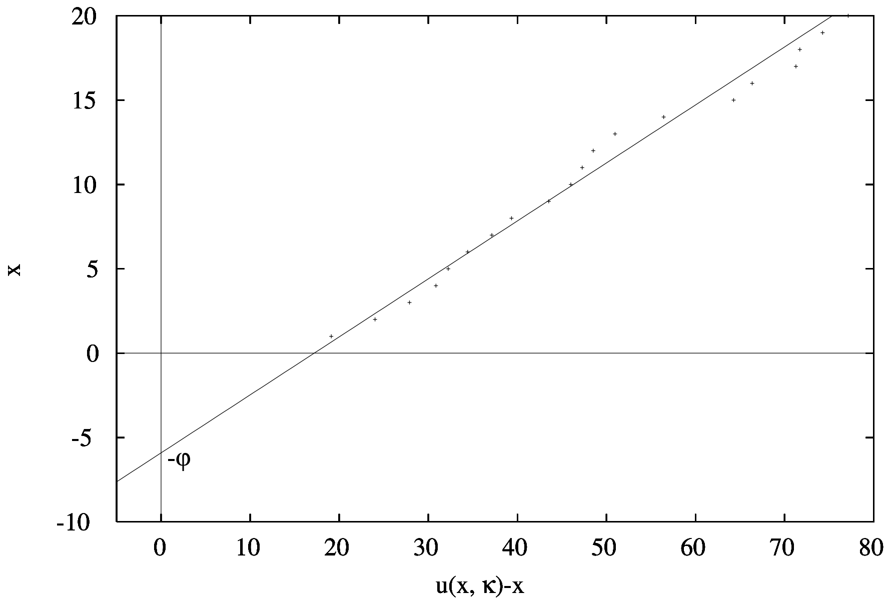

Abbildung 3: Abschätzung der Phasenverschiebung φ

Sind $\hat{a}$ und $\hat{\beta}$ bereits berechnet, kann ein Schätzwert für die Phasenverschiebung φ auch als Mittelwert der Residuen r_i bezüglich der inversen Verteilungsfunktion erhalten werden ($\hat{\alpha} := \hat{\beta} \ln(\hat{a})$):

$$\begin{aligned} r_i &= \left(\frac{\hat{a}}{-ln \hat{F}(x_i)} \right)^{\hat{\beta}} - x_i \\ &= \exp \left(\hat{\alpha} - \hat{\beta} * \ln \ln \frac{1}{\hat{F}(x_i)} \right) - x_i \qquad (5) \end{aligned}$$

Abbildung 4 demonstriert die erreichte Anpassung durch Gegenüberstellung der empirischen Verteilungsfunktion $\hat{F}$ und der errechneten Verteilungsfunktion $\tilde{F}$. Es fällt auf, daß beide Funktionen bis zum Level 99.99% fast deckungsgleich verlaufen; danach entwickeln sich die empirischen Werte seitlich versetzt mit gleicher Krümmung. Dabei könnte es sich um einen reinen Zufallseffekt handeln - die abweichenden Punkte repräsentieren immerhin nur einen Anteil von

[1] Die Konstante κ ist zwar theoretisch frei wählbar, sollte aber weder zu klein (dann starke Streuung der Punktwolke) noch zu groß (dann zu wenige Punkte zur Fixierung der Regressionsgeraden) sein.

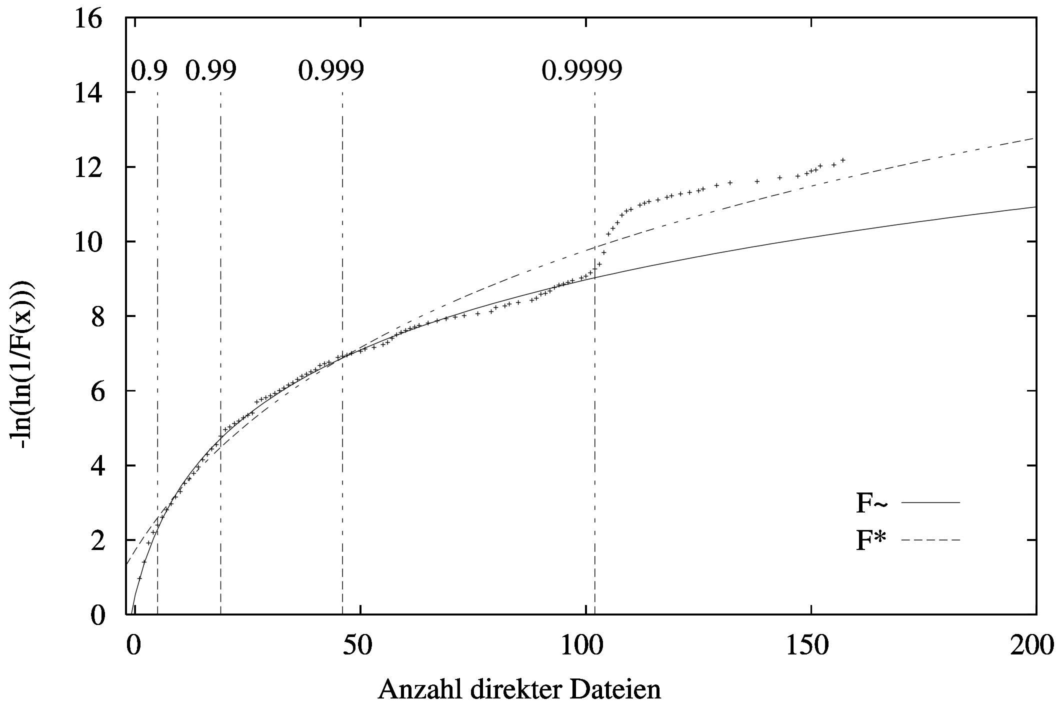

Abbildung 4: Empirische und errechnete Verteilungsfunktion für X_d

weniger als 0.01% der Verteilungsmasse. Eine plausiblere Erklärung für dieses Erscheinungsbild ist allerdings, daß vermutlich einige ungeduldige Nutzer den Übertragungsvorgang bei Files mit sehr vielen Includes vorzeitig abgebrochen haben. Stimmt diese Annahme, müßten extreme Ausreißer in der Stichprobe unterrepräsentiert und die mittleren Werte entsprechend überrepräsentiert sein. Die berechnete Verteilungsfunktion $\tilde{F}$ liefert damit ein etwas zu pessimistisches Bild. Für die Simulation hochausgelasteter Netzwerke kann das sogar ein Vorteil sein. Treten im Simulationsexperiment unter der pessimistischen Annahme von $\tilde{F}$ keine Engpässe auf, dann gilt dies erst recht auch für den Realfall.

Bei der in Abb. 4 als punktierte Linie dargestellten Verteilungsfunktion F^* wurden die Parameter (a, b, φ) unter Weglassen des möglicherweise verfälschten mittleren Abschnittes berechnet. Es ist deutlich zu sehen, daß F^* im hochwahrscheinlichen Bereich nicht so gut angepaßt ist wie $\tilde{F}$, dafür aber die Ausreißer in Dimension und Häufigkeit besser repräsentiert.

Tabelle 3 stellt die berechneten Verteilungsparameter bei separater Betrachtung der von lokalen und der von globalen Hosts angeforderten Dateien und bei gemeinsamer Analyse beider Datengruppen gegenüber.

Überraschenderweise ergab sich die beste Anpassung im Falle der gemeinsamen Analyse aller Daten. Die Unterscheidung in lokale und globale Hosts brachte somit bei der Modellvariablen X_d nicht den erhofften Genauigkeitsgewinn.

Population	Parameter			erreichte Anpassung
	a	b	φ	
alle hosts	$1,73049$	$2,94152$	$5,98998$	++
nur lokale hosts	$1,59652$	$2,75801$	$4,17698$	++
nur globale hosts	$1,77951$	$3,02188$	$6,73181$	++

Tabelle 3: Errechnete Verteilungsparameter aus (1) für die Modellvariable X_d

3.3 Verteilungsgesetz für die Größe direkter Dateien S_d

Die Filegrößen S_d aller im Erhebungszeitraum übertragenen direkten Dateien zeigten ein ähnliches Häufigkeitsbild wie die Population der Beobachtungswerte

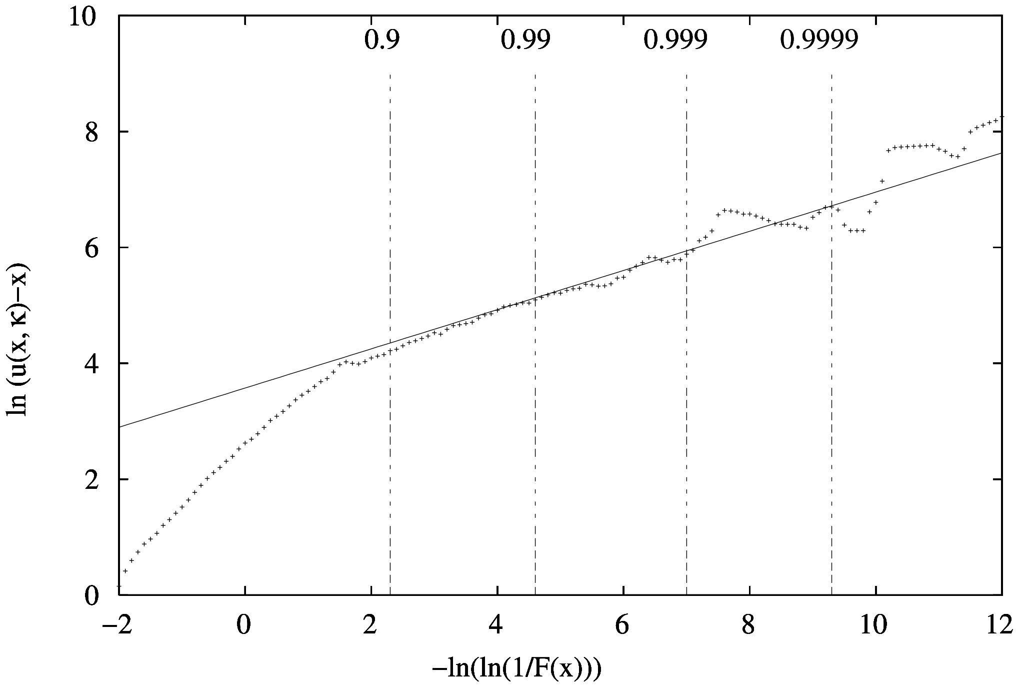

Abbildung 5: Gültigkeit des Fisher-Tippett-Gesetzes für die Filegröße S_d

der Modellvariablen X_d: einer sehr großen Anzahl kleiner Files (z.Bsp. enthielten 90% der Files weniger als 20 kByte) standen einige extrem umfangreiche Dateien

(Maximum: 6.4 MByte) gegenüber. Deshalb wurde zunächst wiederum (1) als Verteilungsfunktion angesetzt. Abbildung 5 zeigt allerdings, daß ein linearer Zusammenhang zwischen den Größen $-lnln\frac{1}{F(x_i)}$ und $ln(u_i - x_i)$ hier nur partiell gegeben ist. Es existiert somit keine Parameterkombination (a, b, φ), die überall - d.h. sowohl für die großen als auch für die kleinen Filegrößen - eine akzeptable Anpassung des Funktionsansatzes (1) an die empirischen Daten liefert. Das Fisher-Tippett-Gesetz gilt offenbar mit sehr guter Näherung ab einer kumulierten Verteilungsmasse von ca. 98% (d.h., für Filegrößen ab 65 kByte, vgl. Abb. 5). Werden die Schätzwerte $\hat{\alpha}$, $\hat{\beta}$ und $\hat{\varphi}$ für die Parameter aus (1) ausschließlich unter Berücksichtigung der großen übertragenen Files berechnet (der Verlauf der empirischen Verteilungsfunktion für $x < 65k$ wird dabei einfach ignoriert), so beschreibt (1) das Verteilungsgesetz für $x \geq 65k$ ausreichend genau.

Für den hochwahrscheinlichen Bereich ist eine Modifizierung des Ansatzes (1)

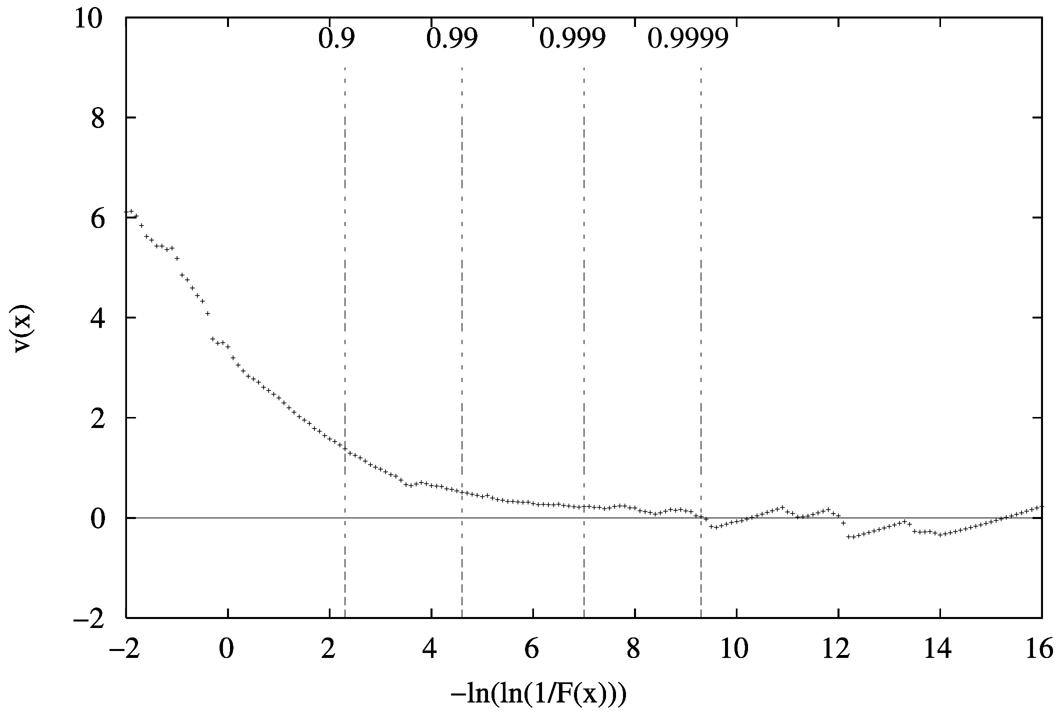

Abbildung 6: v_i nach (6) versus transformierte empirische Verteilungsfunktion

erforderlich. Hierbei zeigte sich, daß nicht die Residuen r_i gemäß (5), sondern die logarithmierten relativen Abweichungen

$$\begin{aligned} v_i &= \ln\left(\frac{e^{\hat{\alpha}-\hat{\beta}*\ln\ln\frac{1}{\hat{F}(x_i)}}}{x_i}\right) \\ &= \left(\alpha - \beta * \ln\ln\frac{1}{\hat{F}(x_i)}\right) - \ln x_i \qquad (6) \end{aligned}$$

eine verwertbare Gesetzmäßigkeit liefern. Abbildung 6 verdeutlicht, daß die gemäß (6) berechneten Werte v_i im Bereich oberhalb $\hat{F}(x_i) > 0.999$ erwartungsgemäß um den Wert 0 schwanken, mit fallender Filegröße aber exponentiell ansteigen. Mit

$$y(x) \quad = \quad -\ln\ln(1/\hat{F}(x)) \tag{7}$$

als Argument läßt sich tatsächlich für den Bereich $y < 7$ (d.h., $\hat{F}(x) < 0.999$ bzw. $x < 310k$) ein funktionaler Zusammenhang der Form

$$v(y) \quad = \quad \gamma \, * \, e^{-\delta \, * \, y} \qquad \gamma > 0, \quad \delta > 0 \tag{8}$$

nachweisen. Schätzwerte für die Parameter γ und δ ergeben sich aus den Parametern der Regressionsgeraden durch die Punktmenge ($-lnln(1/\hat{F}(x_i))$, $ln(v_i)$). Da der Wert der Funktion $y(x)$ für $x \to \infty$ wie $ln(x)$ anwächst, erreicht $v(y(x))$ für große x recht schnell sehr kleine Werte. Die beiden Ansätze (8) und (6) können deshalb zu einem für den gesamten Wertebereich gültigen Funktionsansatz für die inverse Verteilungsfunktion kombiniert werden:

$$x \quad = \quad exp\left(\alpha + \beta \, * \, y - \gamma \, * \, e^{-\delta \, * \, y}\right) \tag{9}$$

Die Generierung von Zufallsrealisierungen für die Modellvariable S_d funktioniert dann nach folgendem einfachen Algorithmus:

1. Erzeuge eine im Intervall [0 , 1] gleichverteilte Zufallszahl u
2. Interpretiere u als Wert der Verteilungsfunktion von S_d und transformiere $y = -\ln\ln(1/u)$
3. Berechne eine Zufallsrealisierung x entsprechend (9)

Die resultierende Verteilungsfunktion F(x) erweist sich damit als Kombination zweier Extremwertgesetze (TypI und TypII) ; sie ist nicht explizit darstellbar.
Bei der Interpretation von (9) muß berücksichtigt werden, daß die ausgewerteten empirischen Daten die *tatsächlich übertragene* Filegröße repräsentieren, welche nicht notwendig mit der Größe der *angeforderten* Datei übereinstimmen muß. Es ist sogar sehr wahrscheinlich, daß ein Teil der Files nur als Fragment angekommen ist. Diese Vermutung wird dadurch gestützt, daß unter allen registrierten übertragenen Files immerhin 11% mit der Länge 0 protokolliert waren.[2]
Die ersten beiden Summanden in (9) sind - dieser Überlegung folgend - als Verteilungsgesetz für die Größe der angeforderten Datei interpretierbar (Fisher-Tippett II analog X_d).

[2]Diese 11 % Verteilungsmasse wurden bei der statistischen Analyse nicht mit berücksichtigt; die ermittelte Verteilungsfunktion ist also strenggenommen die bedingte Verteilung bei stattgefundener (Teil-)Übertragung.

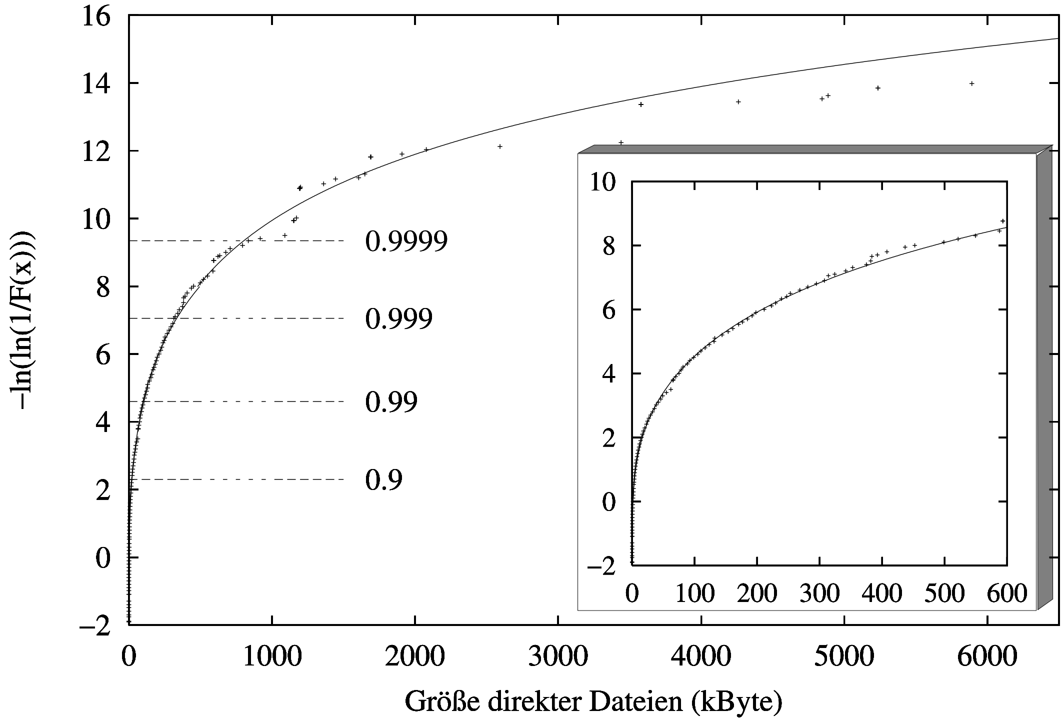

Abbildung 7: Empirische und errechnete Verteilungsfunktion für S_d

Population	Parameter				erreichte Anpassung
	α	β	γ	δ	
S_d	$3,60735$	$0,33819$	$3,30863$	$0,40042$	++
S_h	$2,05245$	$0,41376$	$1,32114$	$0,72379$	+
S_o	$7,07259$	$0,33329$	$4,18771$	$0,32975$	∘

Tabelle 4: Verteilungsparameter nach (9) für die Größe verschiedener Filetypen

Der 3. Term von (9) beschreibt den Beschneidungsvorgang während der Übertragung: ein großer Teil der Files kommt vollständig (im Sinne des Verteilungsgesetzes nur unwesentlich beschnitten) an, ein kleiner Teil als Fragment. Offenbar korrespondieren die Parameter γ und δ mit der Wahrscheinlichkeit und dem Umfang stattfindender Fragmentierungen.
Abbildung 7 demonstriert die mit dem Ansatz (9) erreichte exzellente Anpassung an die empirischen Daten - sowohl für kleine als auch für extrem große Files. Die berechneten Verteilungsparameter aus (9) enthält Tabelle 4.

3.4 Die Größe indirekter Dateien

Ein weiterer wichtiger Modellparameter zur Simulation des Verhaltens eines HTTP-Clienten ist auch die Größe der indirekten Dateien. Das gesamte Verfahren für die Suche nach dem zugrundeliegenden Verteilungsgesetz wurde mit allen Konsequenzen aus Abschnitt 3.3 übernommen.

3.4.1 Die Größe von HTML-Dateien S_h

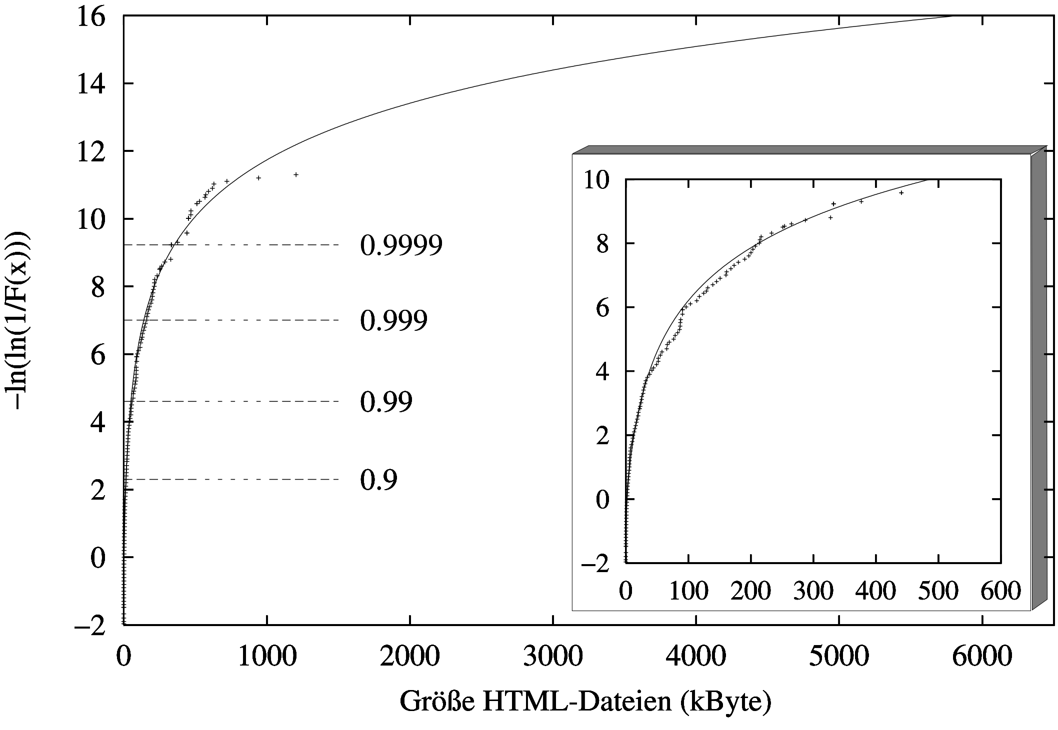

Abbildung 8: Empirische und errechnete Verteilungsfunktion für S_h

Die Analyse der Größe von HTML-Dateien ergab signifikante Unterschiede in den ermittelten Parametern zur Verteilungsfunktion, je nach anfordernden Hosts.

Während globale Hosts durchschnittlich kleinere HTML-Dokumente anforderten, sind es bei lokalen Hosts größere Dokumente. Ursache sind vermutlich absichtliche Übertragungsabrüche durch globale Nutzer, aufgrund mangelnder Beharrlichkeit beim Warten auf die Übertragung größerer Dokumente.

3.4.2 Die Größe von others-Dateien S_o

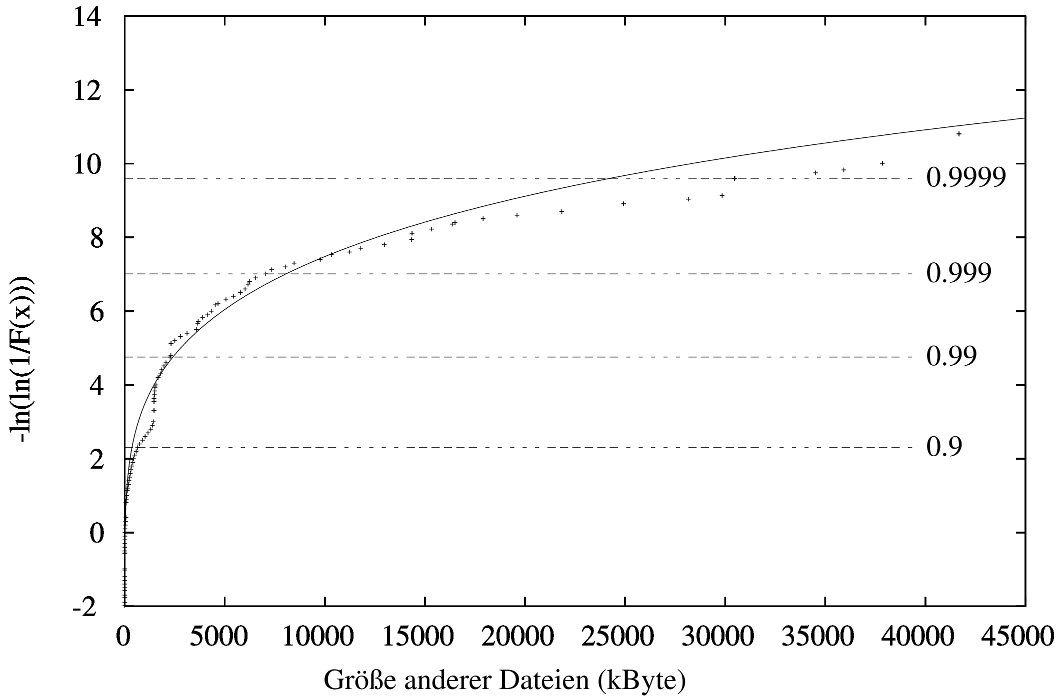

Abbildung 9: Empirische und errechnete Verteilungsfunktion für S_o

Die Übereinstimmung zwischen empirischer und theoretischer Verteilungsfunktion ist hier im hochwahrscheinlichen Bereich (Filesize $x < 1.6$MByte bzw. $F(x) < 0.98$) nicht so gut wie bei den anderen Filetypen. Mögliche Ursache ist einerseits die deutlich geringere Anzahl von Beobachtungswerten bei wesentlich höherer Spannweite (22 Byte bis 46.6 MByte). Weiterhin sprach der gehäufte Abruf von Dokumenten einer bestimmten Größe (ca. 1500 kByte) dafür, daß von vornherein ein glatteres Anwachsen der empirischen Verteilungsfunktion nicht zu erwarten war. Dennoch repräsentiert der Ansatz (9) auch hier die gesamte Spannweite der Verteilung recht gut. In diesem Zusammenhang scheint der Verweis auf einen Vergleich zwischen WWW-Dokumentengröße und UNIX-Filegröße in [CBC95] angebracht.

4 Zusammenfassung

In dieser Arbeit wurden Verteilungsfunktionen gefunden, die eine gute Übereinstimmung mit den untersuchten Parametern des Datenverkehrs von HTTP-Clienten aufweisen und somit als Grundlage für eine Simulation genutzt werden können. Signifikante Unterschiede zwischen einzelnen Monatsauswertungen wurden nicht festgestellt. Von Interesse ist ein Vergleich der erzielten Ergebnisse mit Analysen von Server-Log-Dateien kommerzieller Einrichtungen. Es ist zu beachten, daß die Verteilungsparameter eine dynamische Größe darstellen. So ist davon auszugehen, daß bei Veränderungen im HTML-Standard oder bei Einsatz neuer Kompressionsverfahren für Bilder und Videos Veränderungen eintreten werden. Mögliche Korrelationen zwischen einzelnen Modellparametern wurden nicht betrachtet. Vorstellbar ist insbesondere ein Zusammenhang zwischen der Anzahl X_d direkter Dateien und deren mittlerer Größe. Weiterführende Untersuchungen in dieser Richtung könnten einen Ansatzpunkt für eine Verbesserung des vorgestellten Modells bilden.

Literatur

[AS64] M. Abramowitz and I.A. Segun. *Handbook of Mathematical Functions.* Dover Publications, Inc., 1971/1964. ISBN 0-486-61272-4.

[AW96] M. F. Arlitt and C. L. Williamson. Web Server Workload Characterization : The Search for Invariants. In *Proceedings of SIGMETRICS96*, 1996. http://www.cs.usask.ca/grads/mfal26/final.ps.gz.

[Bob97] C. Bobber. Charakterisierung des HTTP-Datenverkehrs von HTTP-Clienten. Technical report, Technische Universität Chemnitz, 1997. http://rnvs.informatik.tu-chemnitz.de/Mitarbeiter /Thie_sub/open_docs/http_char.ps.gz.

[CBC95] C. R. Cunha, A. Bestavros, and M. E. Crovella. Characteristics of WWW client-based Traces. Technical report, Boston University, 1995. ftp://cs-ftp.bu.edu/techreports/95-010-www-client-traces.ps.Z.

[CDJM91] R. Caáceres, P.B. Danzig, S. Jamin, and D. J. Mitzel. Characteristics of Wide-Area TCP/IP Conversations. Technical report, University of California, 1991. http://netweb.usc.edu/jamin.traffic.

[Gum66] E.J. Gumbel. *Statistics of extremes.* Columbia U.P., 1966.

[Pax94] V. Paxson. Empirically-Derived Analytic Models of Wide-Area TCP Connections. *IEEE/ACM Transactions on Networking*, 2(4):316–336, August 1994.

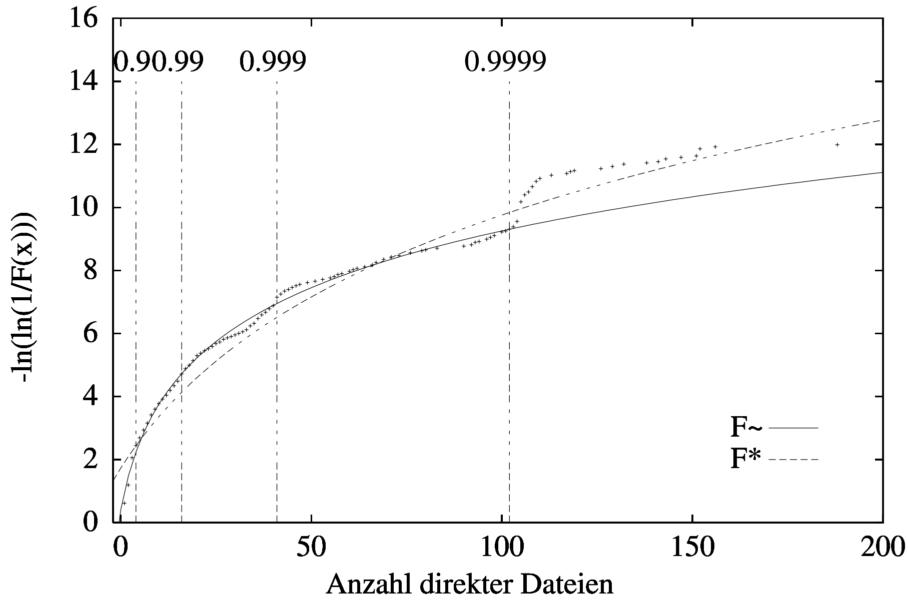

Abbildung 10: Erreichte Anpassung zwischen empirischer und berechneter Verteilungsfunktion für die Anzahl direkter Dateien (lokal)

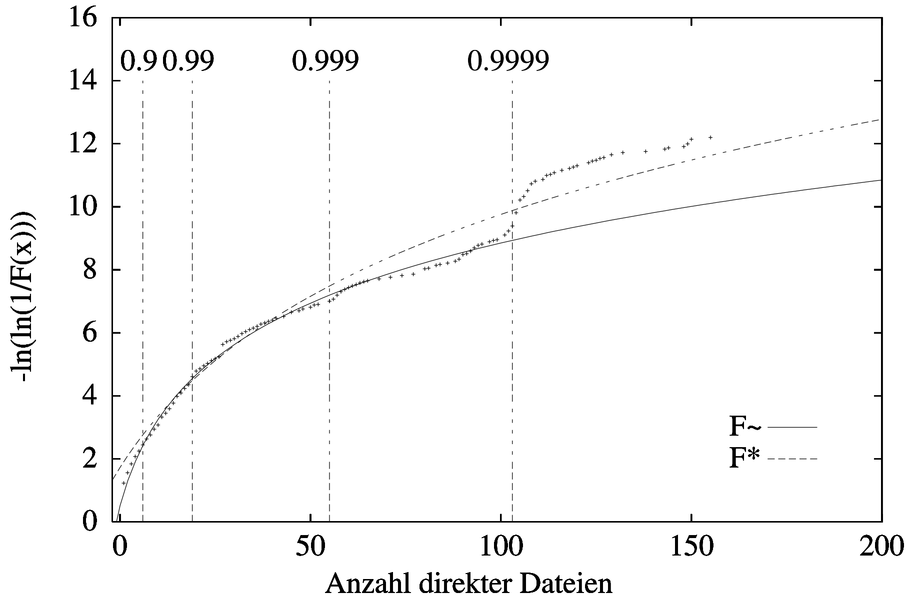

Abbildung 11: Erreichte Anpassung zwischen empirischer und berechneter Verteilungsfunktion für die Anzahl direkter Dateien (global)