Measurement Techniques and Case Studies for the Characterization of Internet Applications

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Universität Dortmund
am Fachbereich Informatik

von

Alexander Klemm

Dortmund
2006

Tag der mündlichen Prüfung:	03.07.2006
Dekan:	Prof. Dr. Bernhard Steffen
Gutachter:	Prof. Dr.-Ing. Christoph Lindemann, Prof. Dr. Heiko Krumm

Abstract

This thesis characterizes the two current killer applications of the Internet: World Wide Web (WWW) and Peer-to-Peer (P2P) file sharing. With the advances in network technology and radical cost reduction for Internet connectivity the Internet grows at an awesome speed in terms of number of users, available content and network traffic. Due to the huge amount of available data, developing algorithms to efficiently locate desired information is a difficult research task. Thus, the characterization of the two most popular Internet applications, which enables the design and evaluation of novel search algorithms, constitutes the two key contributions of this work.

As first contribution, this thesis provides a synthetic workload model for the query behavior of peers in P2P file sharing systems which can be used for evaluating new P2P system designs. Whereas previous work has solely focused on aggregate workload statistics, this thesis presents a characterization of individual peer behavior in a form that can be used for constructing representative synthetic workloads. The characterization is based on a comprehensive 40 days measurement study in the Gnutella P2P file sharing system comprising more than 10 GBytes of trace data. As a key feature, the characterization distinguishes between user behavior and queries that are automatically generated by the client software. The analysis of the measured data exposes heterogeneous behavior that occurs on different days, in different geographical regions or at different periods of the day. Moreover, the consideration of additional correlations among the workload measures allows the generation of realistic workloads.

As second contribution, this thesis characterizes and models the structural properties of German Web sites for enabling their automated classification. These structural properties encompass the size, the organization, the composition of URLs, and the link structure of Web sites. In fact, the approach is independent of the content of Web pages. Opposed to previous work, this thesis characterizes structural properties of entire Web sites instead of individual Web pages. The measurement study is based upon more than 2,300 Web sites comprising 11 million crawled pages categorized into five major classes: *Brochure*, *Listing*, *Blog*, *Institution*, and *Personal*. As a key insight which can be exploited for improving Internet search engines and Web directories, this thesis reveals significant correlations between the structural properties and the class of a Web site.

Acknowledgement

Many people have their share in the achievement of this work (in one way or the other). A special thank goes to my thesis advisor Prof. Dr.-Ing. Christoph Lindemann who introduced me to scientific working methods and who was always available in case of questions. Some of this thesis' research results were developed in co-operation with Dr. Oliver Waldhorst and Lars Littig. I would like to thank them for numerous fruitful discussions. Moreover, I would like to thank all colleagues of the *Computer Systems and Performance Evaluation* group for a motivating and relaxing work atmosphere. This friendly ambience has tremendously eased the completion of this work.

My particular thanks go to my mother Carmen, my brother Georg and my sister-in-law Nadine. They offered full family support which enabled the completion of this work in the first place. I dedicate this work to my father Klaus who always believed in me.

Danksagung

Viele Personen haben auf die eine oder andere Weise Anteil am Gelingen dieser Arbeit. Mein besonderer Dank geht an meinen Doktorvater Prof. Dr.-Ing. Christoph Lindemann, der mich an wissenschaftliche Arbeitsmethoden heranführte und stets Ansprechpartner für alle Fragen war. Einige Forschungsergebnisse dieser Arbeit entstanden in Zusammenarbeit mit Dr. Oliver Waldhorst und Lars Littig. Beiden möchte ich für die vielen fruchtbaren Diskussionen danken. Weiterhin bedanke ich mich bei allen Kollegen des *Fachgebiets Rechnersysteme und Leistungsbewertung* für eine motivierende und entspannte Arbeitsatmosphäre. Das freundschaftliche Umfeld hat die Fertigstellung dieser Arbeit ungemein erleichtert.

Mein spezieller Dank gebührt meiner Mutter Carmen, meinem Bruder Georg und meiner Schwägering Nadine. Sie boten den kompromisslosen familieren Rückhalt, der die Fertigstellung dieser Arbeit erst ermöglichte. Ich widme diese Arbeit meinem Vater Klaus, der immer an mich glaubte.

Contents

Chapter 1

Introduction

1.1 Motivation

In the relative short time span of its existence, the Internet has become one of the most important media for information exchange and personal communication. Today, the Internet provides the basis for an increasing number of Internet applications like telephony, video on demand, gaming, Web browsing, international commerce, file sharing, online banking, etc. The growth of Internet usage is expressively demonstrated by the fact that the number of Internet users increased from 559 millions in 2000 to 1,018 millions in 2005 [Min06]. In the same period, the number of active Web sites increased from about 7.6 millions to about 34 millions [Net06].

Although available bandwidth increases, there is widespread agreement that the increasing network traffic can only be handled by an intelligent optimization of resource usage. To achieve this challenging goal numerous efforts have been started in the last few years to measure and characterize the Internet [Coo06]. The characterization of the Internet is essential to get insight in how the Internet behaves. In particular, researchers are interested in the dynamic structure of the Internet, network traffic characteristics, user behavior, and application specific aspects. Furthermore, the characterization of Internet properties enables the creation of stochastic models which reflect the measured Internet properties. By means of these models, novel technologies and protocols can be developed and evaluated in order to achieve scalability and improve system performance.

Regarding the transferred data traffic, the most popular Internet applications constitute Web browsing and Peer-to-Peer (P2P) file sharing. Both applications combined make up more than 80% of the transferred data in the Internet backbone [FML$^+$03]. Consequently, these applications are subject to current research activities. Due to the innumerable number of pages in the Web and the immense amount of files shared in P2P file sharing systems, efficiently locating desired information is a difficult task in both applications. Therefore, this thesis deals with the query behavior of P2P file sharing users in order to enable the development of improved algorithms for file search. As a second contribution, this thesis characterizes Web sites in order to improve the search experience made with Internet search engines.

P2P File Sharing Systems

P2P file sharing systems provide algorithms and protocols for sharing files between individual users of the system (*peers*). They invented a totally new communication pattern in the Internet, the *P2P paradigm*. In contrast to traditional client-server communication, where many clients communicate with dedicated central servers, in P2P systems involved entities are both client and server. These peers not only download files from other hosts in the role of clients but also act as servers since they share locally stored files for download by other peers.

These systems became popular with the advent of the Napster system which was designed for sharing MP3 music files between users without the bottleneck of a central storage server. Although the actual transfer of files in the Napster system was processed directly from peer to peer, the location information of files, i.e. which file is shared by which peer, was stored on a central index server. Thus, this type of P2P system is denoted as *centralized P2P system* [LCC$^+$02]. Current representatives of centralized P2P file sharing systems include eDonkey [Met05] and BitTorrent [Bit04].

Not at least because of copyright issues which terminated the Napster system, novel approaches with an entire decentralized structure gain increasing popularity. These *decentralized P2P systems* establish TCP connections between peers in order to build an application-level *overlay network* which is used for locating files to download. This overlay network can either be *structured*, as in systems based on distributed hash tables (DHT) like CAN [RFH$^+$01] and CHORD [SMK$^+$01], or *unstructured* as in FastTrack [Sha04] and Gnutella [Gnu04]. In structured P2P systems the overlay construction is highly controlled and the location (or the information about the location) of resources is uniquely determined by the identifier of the resource. Since these systems do not support keyword-based search and their performance in highly dynamic environments is unknown, there are currently no widely deployed P2P file sharing systems based on DHTs [CRB+03]. On the contrary, in unstructured P2P systems, the location of resources is determined by a search process which generally incorporates "flooding" the entire or parts of the overlay network. These P2P systems support keyword-based search and are resilient to frequent leaves, joins, and topology changes.

Designing a search protocol for a P2P file sharing system, regardless of whether it is structured, unstructured, or following any other approach, requires the evaluation of different design alternatives. In early stages of the design process, analytic models can support design decisions by providing aggregate measures of protocol performance. However, later design stages require detailed simulation studies or even field studies based on software prototypes. Such performance evaluations require both a detailed model of the considered system as well as a detailed workload model to mimic the load that the system has to bear during operation. An important aspect of a detailed workload model constitutes a user's active behavior, i.e. the generation of queries. For example, Chawathe et al. [CRB+03] use simulations of client query behavior to evaluate a new overlay network architecture and a new biased random walk search protocol, whereas Ge et al. [GFJ+03] use an analytic model of query behavior to compare alternative directory architectures and search protocols.

Recent studies have provided important partial characterizations of peer behavior, including aggregate distributions of session durations, time between downloads, query and file popularity, requested file sizes, and measured bandwidth between the peer and the Internet at large [BSV03], [GDS+03], [SGD+02], [SGG02], [SW02], [Sri01], [YGM01]. Several of these studies consider the impact of time of day or another specific correlation between the measured parameters. However, previous workload studies have two significant drawbacks for constructing realistic synthetic workloads: (1) they are incomplete with respect to correlations among the workload measures, and (2) they include aggregate measures that obscure heterogeneous behavior across different classes of peers or across different periods of time. In Chapter 3, this thesis presents a detailed workload model for the query behavior of P2P file sharing users which captures all key correlations and allows the synthetic generation of query workload for individual users.

World Wide Web

While a key research problem for P2P file sharing systems is to improve the efficiency of the search algorithm, enhancing the quality of search results in the Web is another current research topic. The task of finding relevant information in the Web is in general processed by Internet search engines like Google [Goo05] or MSN Search [Mic05] or by Web directories like Yahoo [Yah05] and DMOZ [Net05]. Enhancing the functionality and/or operation of search engines and directory services is crucial for coping with the immense growth of the Web. The four key components of Internet search engines constitute the crawling strategy, the refreshing strategy, the ranking method, as well as a spam detection method [ACGM$^+$01]. Crawling strategies specify which of the known Web pages shall be crawled and indexed by the search engine. Refreshing strategies determine how often the search engine shall update the Web pages of the index. Ranking methods determine the order in which the search engine displays the URLs of Web pages matching a search query. Spam detection methods identify collections of Web pages or entire Web sites with useless content, whose only purpose lies in fooling the ranking methods of popular Internet search engines.

Today's leading Internet search engines already distinguish in their operation between different top-level domains. For example, the order in which the pages matching a search query are displayed by the search engine Google [Goo05] is different in `www.google.com` and `www.google.de`. Further knowledge, such as some classification of individual Web sites within a particular top-level domain, will be extremely valuable for improving the capabilities of search engines. In fact, the coarse classification of individual Web sites will allow the improvement of each of the four key components of search engines mentioned above. First, the coarse classification of Web sites can support *focused crawling*, i.e. it provides means to crawl only sites which belong to a specific class, thus enabling specialized search, e.g., Web log search. Second, as shown in [CGM00] and [FMNW03], the evolution of Web pages correlates to the class of the corresponding site. For example, pages in the `.com` domain change more frequently than in the `.edu` domain. Thus, knowing a site's class (not only the top-level domain) can certainly improve the refreshing strategy. Third, ranking algorithms can benefit from site classification by tagging search results with the class of the corresponding Web site or grouping them by presenting the top ten results for each class. Furthermore, personalized ranking could be performed by favoring results from a certain class of interest.

Fourth, the identification of spam sites would enable their filtering and, thus, improve the overall search experience. Moreover, the automatic classification of Web sites can ease the manual maintenance of Web directories by an automatic preselection of categories.

Since a very coarse classification of Web sites as commercial, organizational, or educational can be performed by considering their top-level domain, e.g. `.com`, `.org`, or `.edu`, it is in part trivial for some Web sites which are often physically located in the US. But as online shops, information portals, and companies share the top-level domain `.com`, the classification becomes inaccurate. Furthermore, Web sites of various classes reside in the same top-level domain in countries other than the US, e.g. `.ch`, `.de`, or `.fr`.

The main disadvantage of the approaches presented in scientific literature, e.g. [EKS02], [KS04], and [THG$^+$03], is that they are content-based, i.e. they rely on the occurrence of specific terms which are related to different topics. These approaches inherently depend on the language used in the pages. Thus, a content-based classifier must be trained separately for each language. Furthermore, those classifiers are very expensive in terms of computational costs because the feature-space, i.e. the number of different terms considered for the classification, is in general very large. Chapter 4 presents a comprehensive characterization of German Web sites which reveals that there are significant differences in the structural properties of Web sites belonging to different classes. This characterization builds the basis for novel content-independent classification approaches which are solely based on a small number of structural properties.

1.2 Summary of Contributions of this Thesis

This thesis provides characterizations of the two most popular applications in today's Internet. Thus, its contribution is two-fold. For improving the search efficiency of P2P file sharing systems, the characterization of the query behavior of individual peers provides means for the development and evaluation of novel system designs. Furthermore, a detailed characterization of the structural properties of Web sites builds the basis for enhancing the quality of search results in Internet search engines and Web directories. This section summarizes the main results for both applications.

Characterization of the Query Behavior in Peer-to-Peer File Sharing Systems

As first contribution, this thesis presents a synthetic workload model for the query behavior in P2P file sharing systems. In a passive measurement study conducted in the Gnutella P2P file sharing system [Gnu04] over a period of 40 days, more than four million connected sessions are recorded. Based on this measurement study, this thesis analyzes the query behavior of P2P file sharing peers. Opposed to previous work which provides various aggregate workload statistics, we characterize peer behavior in a form that can be used for constructing representative synthetic workloads for evaluating novel P2P system designs. Therefore, the query behavior of individual peers is analyzed by considering only query messages for which the issuer can be uniquely identified. Furthermore, queries that are specific to the Gnutella system are eliminated, such as automated re-queries that are issued by some client implementations. Thus, the obtained results can be applied to P2P file sharing systems other than Gnutella as well.

The characterization of the query behavior considers all workload measures which are necessary for building a synthetic workload model. These measures include the fraction of connected sessions that are completely passive (i.e. no queries are issued), the duration of such sessions, and for each active session, the number of queries issued, the time until sending the first query, the time between sending two subsequent queries, the time after sending the last query, and the query popularity. To reveal environmental influences on user behavior, the characterization examines the correlation to the geographical region of the issuer and the time of day for each workload measure. The results show that there are significant differences in the user behavior of peers in different geographical regions (i.e. Asia, Europe, or North America) and at different times of the day. Moreover, the characterization exposes additional correlations among the workload measures which have to be considered for generating realistic workload.

Based on the characterization of the individual workload measures, this thesis develops a synthetic workload model for the query behavior of individual peers. This workload model is composed of probability distribution functions which are fitted to the measured data and captures all key correlations in the measures in the form of conditional distributions. Moreover, this thesis provides all required distribution functions and parameters, such that the workload model can be directly applied in performance models for P2P file sharing systems.

Characterization of Web Sites by their Structural Properties

The second contribution of this thesis comprises the measurement-based characterization of Web sites in terms of their structural properties. The structural properties reflect the size, the organization, the composition of URLs, and the link structure of Web sites. Compared with previous work, which either just considers the link structure as [BKM$^+$00], [DKM$^+$02] or the change ratio of content of Web pages as [CGM00], [NCO04], [FMNW03], the proposed set of measures provides a comprehensive characterization of entire Web sites rather than just considering individual pages. Since these measures entirely rely on structural properties of Web sites, they can be derived without inspecting the content of Web pages and thus are language-independent.

This thesis presents results of a comprehensive Web measurement study of Web sites belonging to the twelve major *categories* of the German Web. These categories include, amongst others, academic institutions, Web logs, online forums, information portals, online shops, small and medium-sized enterprises, and private homepages. Overall, more than 2,300 Web sites with more than 11 million Web pages are crawled. The analysis of the measured data shows that Web sites belonging to specific categories are similar in terms of their structural properties, whereas there are significant differences to Web sites belonging to other categories. Thus, similar categories of Web sites can be grouped into *classes* which significantly differ in their structural properties. A key observation of the analysis is that Web site categories grouped into the same class are not only similar with respect to structural properties but are also functionally related. For example, Web sites of the categories *online shop* and *information portal* are grouped into the same class *Listing* because they have both similar structural properties and similar function in that they list numerous items.

Based on this observation, we present a detailed characterization of Web sites belonging to the classes Brochure, Listing, Blog, Institution, and Personal with the goal of identifying structural properties of Web sites which allow their automated coarse classification. The characterization reveals significant correlations between the structural properties and the class of a Web site as well as certain correlations among the structural properties. Furthermore, this thesis provides probability distributions and fitted parameters for direct application in classifiers as, e.g., the naïve Bayes classifier [DHS01].

1.3 Publications Making up this Thesis

The characterization of the query behavior in P2P file sharing systems presented in Chapter 3 has been published in [KLVW04] and with a different focus in [KLW04b]. Both papers have been co-authored with Oliver Waldhorst. In [KLVW04] and [KLW04b], the author of this thesis developed the measurement setup and the framework for characterizing user behavior. Oliver Waldhorst conducted experiments to confirm the representativeness of the measured data and supported the analytical modeling. The paper [KLVW04] was also co-authored by Mary Vernon who supported the completion of the paper by discussion of the measurement results.

The characterization of Web sites by their structural properties presented in Chapter 4 has been jointly performed with Lars Littig. The author analyzed the measured data, characterized the structural properties of Web sites and derived model distributions for each measure. Lars Littig is developing a classifier which exploits the measurement results for the automated classification of Web sites.

Furthermore, the author of this thesis co-authored several publications which are not part of this thesis. [KLL01] provides a synthetic workload model for UMTS traffic and is co-authored with Marco Lohmann. In this paper, the author measured and characterized the Internet traffic of dial-in users, whereas Marco Lohmann developed a mathematical framework to derive effective traffic models. The papers [KLL02] and [KLL03] are also co-authored with Marco Lohmann and provide a novel approach for modeling IP traffic. These papers are part of Marco Lohmann's Ph.D. thesis [Loh04]. The author contributed to [KLL02] and [KLL03] by measuring and characterizing the Internet traffic behavior of dial-in Internet users. The analysis of the measured data served as comparative modeling approach for an analytic traffic model based on the Batch Markovian Arrival Process. Marco Lohmann developed a novel parameter estimation method for this analytic traffic model.

In another paper co-authored with Sherif ElRakabawy, the author co-developed a novel variant of the Transmission Control Protocol (TCP) with significantly improved performance in Mobile Ad Hoc Networks (MANET) [EKL05]. In this paper, the author identified the variation in the round-trip times (RTT) as effective measure for contention in MANET. Sherif ElRakabawy developed a rate-based formula for appropriately pacing the TCP sender and conducted the simulation experiments.

The papers [KLW03] and [KLW04a] are co-authored with Oliver Waldhorst and part of his Ph.D. thesis [Wal05]. In [KLW03], the author developed the transport protocol for the ORION system while Oliver Waldhorst designed the search algorithm. In [KLW04a], the author showed at the example of ORION that cross-layer information transfer can improve the performance of P2P systems in MANET. Oliver Waldhorst discussed the shortcomings of off-the-shelf P2P systems in such environments and presented PDI as a possible solution based on epidemic data dissemination.

The publications [LTK$^+$00] and [LTK$^+$02] are part of the Ph.D. thesis of Axel Thümmler [Thü03] and deal with the performance analysis of time-enhanced UML diagrams. In these papers, the author developed the algorithm for converting time-enhanced UML diagrams into stochastic Petri nets which are the basis for performance analysis by means of stochastic processes.

1.4 Thesis Outline

This thesis is organized as follows. Chapter 2 compiles and recalls basic techniques for the measurement and characterization of Internet applications. Therefore, it provides the methodological background for the measurement-based characterizations presented in Chapters 3 and 4. It summarizes important aspects for conducting meaningful and representative measurements in Section 2.1 and outlines three approaches for analyzing correlations in Section 2.2. An introduction to characterizing measured data with probability distribution functions is given in Section 2.3.

Chapter 3 characterizes the query workload in P2P file sharing systems based on measurements in the Gnutella network. After recalling the basic functionality of the Gnutella protocol and discussing previous work in Sections 3.1 and 3.2, respectively, Section 3.3 outlines the measurement methodology tailored to the Gnutella overlay network and shows the representativeness of the measured data. A detailed workload model for the query behavior in P2P file sharing systems based on the characterization of the measured data is presented in Section 3.4. This section in particular includes a complete set of distribution functions and parameters which reflect the query behavior of individual peers and capture all significant correlations. Furthermore, the major results of the characterization are summarized in Section 3.5

The characterization of structural properties of Web sites is the subject of Chapter 4. Section 4.1 relates previous characterization results for the Web to the characterization of Web site structure presented in this chapter. Afterwards, Section 4.2 states the specific methodology employed for the sound measurement of Web sites belonging to the major categories of the German Web and provides statistics of the measured data. Section 4.3 characterizes the structural properties of the measured Web sites and identifies substantial differences between Web sites of different classes. To provide further insight into the different structures of Web sites, Section 4.4 analyzes the correlation structure between the structural properties. After outlining the application of the presented characterization for the automated classification of Web sites in Section 4.5, the main results are summarized in Section 4.6. Finally, Chapter 5 summarizes the results of this thesis and gives some concluding remarks.

Chapter 2

Methodology for Measurement and Characterization

The characterization of specific aspects of the Internet is commonly based on measurements which allow to deduce the considered user behavior, traffic pattern or structural properties. Since all conclusions of the characterization are drawn from these measurements, it is extremely important to invest sufficient effort in its conception and execution in order to gain meaningful results. Furthermore, the characterization of the measured data requires sound methodology to capture all important aspects of the system under examination. Therefore, this chapter summarizes the methodology for measuring and characterizing Internet applications which is applied for the case studies presented in Chapters 3 and 4. First, it gives some insight in how to conduct measurement studies such that the obtained data is representative for the considered system in general. Depending on the goal of the measurement study it may be necessary to identify potential correlations between individual measures. Thus, this chapter provides the background and practical guidelines to analyze the correlation structure between such measures. Finally, for characterizing and modeling the measured data it recalls the fundamentals of finding appropriate probability distributions and corresponding parameters which closely capture the statistical properties of the measured data. Note that a detailed discussion of measurement techniques, which are tailored to the Internet applications P2P file sharing and World Wide Web, is given in Chapters 3 and 4, respectively.

2.1 Measurement Methodology for Obtaining Representative Data

2.1.1 Measurement Setup

One fundamental prerequisite for conducting meaningful measurements in the Internet is to design a sophisticated measurement setup. There are mainly two fundamental approaches for Internet measurements: *passive measurements* which observe network traffic without actively triggering some action in the network and *active measurements* which actively send out probe packets to measure characteristics of the considered system. Passive measurements are often used for measuring user behavior, for example the usage pattern of specific applications or the bandwidth usage at specific network locations. This approach has the advantage that the considered system is not influenced by the measurements, thus the measurement results are not biased by the measurement itself. Since in active measurements the executor of the measurement actively triggers some reaction of the system, this approach comprises a more controlled measurement environment because it does not rely on third party activities. Therefore, active measurements are often used for observing system characteristics such as network latency or bandwidth between two network locations or response times of an Internet server. Here, additional precautions have to be taken to assure that the measurement does not bias the results in a harmful manner. E.g. consider a measurement study which aims at characterizing the average delay of data packets between two network locations A and B by periodically sending probe packets from location A to location B. Then, it is crucial for meaningful results that the network load induced by the measurements does not bias the observed latencies. I.e. the time interval between sending two successive probe packets must be large enough so that the synthetic network load caused by the probe packets is negligible.

In fact, for both active and passive measurements an imprudent measurement setup may have considerable influence on the measurement results and even result in severe measurement errors. Thus, a number of issues has to be considered when designing the measurement study. This section concentrates on important aspects of the design process for the measurement studies presented in Chapter 3 and Chapter 4 of this thesis. For a more general discussion of common pitfalls related to Internet measurements and strategies to overcome those, we refer to [Pax04].

The two major concerns which have to be addressed in the conception of the measurement studies are related to *accuracy* and *representativeness*. Accuracy denotes the correctness and precision of the measured data. That is, we have to assure that the measured data truly reflects the data which we believe to measure. For example, packet filters, traffic shapers or firewalls may render the measurement study useless, if the influence of these network devices is not considered in the measurement setup. E.g., for the passive measurement study of P2P traffic conducted in the overlay network of a P2P file sharing application presented in Chapter 3 it was crucial that the Center for Communication and Information-processing (Hochschulrechenzentrum) agreed to disable the traffic shaper for P2P traffic for the measurement node. Otherwise, the measurement results would have been biased by the traffic shaper in terms of much less traffic load and an unpredictable dropping of P2P messages. Furthermore, the time-related measures of the measurement study strongly depend on a sufficient accurate clock. I.e. for accurate measurements of packet interarrival times at a 1 GBit/s interface the usage of the unsynchronized hardware clock of an off-the-shelf PC is very likely not sufficient. And even for time-related measurements with time units of 1 second, appropriate precautions may be necessary in order to synchronize the clock at the measurement node with other nodes. In the measurement study of P2P workload we solved this issue by synchronizing the clock of the measurement node using the network time protocol (NTP) [Mil92].

The representativeness of the measured data is the second fundamental basis of meaningful measurement studies. It guarantees that the conclusions deduced from the measured data are also true for the considered system in general. That is, for drawing conclusions for the Internet in general, the measured data must not be biased by environmental influences like the network location of the measurement node or the time of measurements. For example, a measurement study of network traffic conducted at daytime hours can not serve as base for the characterization of network traffic at night hours, because user behavior typically follows some diurnal pattern. Similarly, conclusions drawn from the log file of a Web server located in Europe do not necessarily hold for Web servers in Africa, due to different cultural or legal environments. Solutions for these issues are to either expand the measurement to include sufficient data (i.e. both daytime and nighttime or several geographical regions) or to limit the conclusions to those areas for which the measured data is indeed representative. A more detailed discussion on both accuracy and representativeness of the measurement studies presented in this thesis is given in Chapter 3 for the characterization of P2P file sharing workload and in Chapter 4 for the characterization of the structural properties of Web sites, respectively.

2.1.2 Choosing Significant Measures

Besides identifying the proper location and environment for the measurement study as outlined above, the design of the measurement setup involves the choice of appropriate measures. Obviously, most measures are defined in a straightforward manner by the goal of the measurement study. Consider e.g. a measurement study analyzing the response time of a Web server. Depending on the notion of "response time" the measure to capture might be the time between sending out the request and receiving the corresponding response or the correct measure might be the time between the arrival of the request at the server and the time of sending the corresponding response. Whereas the former definition incorporates network delays and processing time at the server, the latter incorporates processing time only. Thus, it is crucial to accurately specify, what measures are meant to be analyzed and rethink if these are not exactly the measures which actually are observed.

In many workload studies it is not possible to directly observe all required measures due to different reasons. Therefore, it may be necessary to include additional measures in the measurement study which in combination allow to extract the desired data. E.g., if in the example above the average processing time of the Web server has to be measured but the server cannot be instrument for direct measurements, this information could be extracted by measuring the response time including network delays and additionally measuring network delays separately (e.g. by using the TCP timestamp option) in order to extract the processing time indirectly.

Applied to the field of P2P file sharing workload this means that a direct measurement of user behavior can only be performed by observing the interaction between the user and the P2P file sharing client. Therefore, the client has to be instrumented in order to log the user behavior for later processing. Obviously, in such an approach finding an appropriate set of users, which agree to use the instrumented client, is difficult. Furthermore, it would be almost impossible to ensure that this user set is representative for all users. Thus, the P2P workload measurement study presented in Chapter 3 indirectly measures the user behavior by passively tracing the generated network messages. This task incorporates additional provisions for identifying the issuer of a message and for distinguishing user behavior from automated client behavior. These provisions, which are crucial for extracting user behavior from the observed network messages, are discussed in detail in Section 3.3.

However, as stated above, additional attention should be paid on the accuracy of the measured data. Including additional measures or meta-data enables cross-checking of the measured data and thus provides means for identifying possible measurement errors. E.g., storing the status code of the HTTP request for each Web page allows to identify broken links, misconfigured Web servers etc. This information can be used for sanitizing the measured data to only include authentic data. Furthermore, additional information helps in interpreting specific characteristics, which could not be explained otherwise. For example storing the software client version of the measured peers in the measurement study of Chapter 3 enabled the identification of peers using misbehaving client software.

2.1.3 Outlier Analysis

An important tool for identifying possible measurement errors is outlier analysis. This method examines the measured data in order to find exceptional, uncommon or unexpected values. That is, looking for values which are unusual low or high or which are measured extremely frequent or seldom, provides hints for possible measurement errors. Furthermore, these outliers, when not caused by measurement errors, may well indicate some system behavior not expected during the design of the measurement setup and thus provide additional insight into the examined system. For example, in the measurement study for characterizing the query behavior of P2P file-sharing users we observed an unusual large fraction of queries from the same peer with interarrival times of 10 seconds. A closer investigation of this phenomenon revealed that this behavior is caused by an automatic re-query feature of the client software instead of regular user behavior. Similarly, in the measurement study of the structural properties of Web sites we observed that some Web sites have extremely large number of slashes, i.e. 100 and more, in the URLs. These outlier values are caused by spamming Web servers which try to improve search engine ranking by pretending large Web sites with many pages (*cloaking* [GGM05]). Since the observed data does not correspond to their real structure, these Web sites have to be filtered out.

The advantage of outlier analysis is that many measurement errors and noteworthy system behaviors manifest in those outliers and its simple application. Outliers can easily be identified by graphically analyzing the measured data. In graphs showing the distributions of the measured data they appear as values which are unexpected small or large or as significant jumps in the curves either to a very frequent value or a very infrequent value. The identification of outliers helps to analyze the measured data in more detail at significant points. This examination has to reveal the cause of the outlier. Depending on this cause for the outlier there are different options for a further processing of the data. In case the outlier is due to a measurement error there may be the possibility to sanitize the measured data by discarding erroneous data. Otherwise, as in the case of a misconception, the measurement setup has to be improved and the measurement redone. In case of an unexpected but correct system behavior the identification of outliers may provide deeper understanding of the system.

2.2 Analyzing the Correlation Structure between Measures

After performing the measurement study and ensuring that the measured data is correct and representative, a further step in characterizing Internet applications is to analyze correlations. Correlations among the individual measures making up a measurement study reveal significant insight in how the system behaves. Furthermore, if the study aims at developing a model of the examined system, the consideration of potential correlations is crucial for developing a representative model which reflects the real system.

In this context, correlation between two measures denotes the dependence of one measure on the other measure. Considering the measures of the measurement study as random variables and the measured values as samples from a random experiment, two measures X and Y are correlated if and only if the corresponding random variables are *not* stochastically independent. X correlates to Y if the probability distribution of X depends on the probability distribution of Y, i.e.

$P\{X|Y\} \neq P\{X\}$.

Since the Internet is a worldwide medium which spans several geographical, cultural, and economic regions and time zones there may well be differences in system behavior between different parts of the world. Therefore, results obtained from measurements conducted in a specific location of the Internet (i.e. part of the world) must not be true for the entire Internet. And likewise, results obtained from measurements in the entire Internet must not be true for a specific part of the world. Thus, to find out if the distribution of a specific measure depends on environmental settings like geographical region or time-of-day, the correlation between the considered measure and the geographical region and time-of-day has to be analyzed. With this correlation analysis it is possible to figure out if the geographical region or the time-of-day effects the system behavior.

Besides the correlation to environmental settings, the correlation between different measures plays also an important role for the characterization of Internet properties. As very simple example, consider a study which measures the number of download requests of PDF documents from a library's Web server and the corresponding file sizes. The number of download requests may well correlate with the file size, e.g. such that small documents like separate chapters of a book are downloaded more often than large PDF documents containing the entire book. Thus, a workload model not incorporating this correlation would clearly generate workloads which does not reflect the reality. In fact, a correct workload model has to incorporate the correlation structure between the measures, i.e. in the example it would have to provide conditional distributions for the number of download requests depending on the corresponding file sizes. In the following we provide two graphical and one mathematical technique for the identification and quantification of correlations between different measures.

2.2.1 Conditional Distributions

The most intuitive method for identifying correlations between two measures X and Y is to plot the probability distribution of X conditioned on Y. Since measurement studies can in general only observe a finite number of events, in our measurement studies we approximate the real probability distribution F_X of the random variable X producing events x observed in the measurement by the *experimental distribution.* The experimental distribution is given by the relative

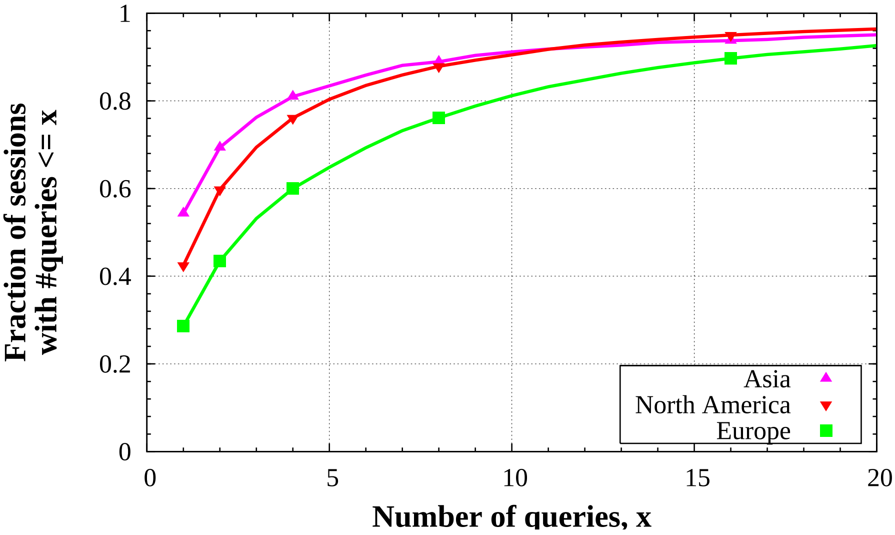


Figure 2.1: Identification of correlations using conditional distributions

frequencies of the observed events, i.e.

$$F_X(t) = P\{X \leq t\} \approx \frac{||(x \in \mathcal{S} | x \leq t)||}{||\mathcal{S}||},$$

where $\mathcal{S}$ is the set of values observed in the measurement. Let F_X and F_Y be the experimental distributions of measures X and Y, respectively. And let X taking values from $\mathcal{S}$ and Y taking values from $\mathcal{T}$. Note that both $\mathcal{S}$ and $\mathcal{T}$ are finite sets, since the measurement study captures a finite amount of data. Then the distribution of X conditioned on Y is given by $F_{X|Y}(x|y) = P\{X \leq x | Y = y\}$ with $x \in \mathcal{S}$ for each $y \in \mathcal{T}$. That is, for each observed value y of the random variable Y the conditional distribution of X is plotted by only considering those observed values of X for which the corresponding value of Y equals to y.

As an illustrative example, consider the number of queries issued per session in a P2P file-sharing system. Without explaining this experiment in detail (this is done in Section 3.4.5) Figure 2.1 plots the distributions of this measure conditioned on the geographical region of the issuer of the corresponding queries. The figure intuitively shows that the number of queries issued in a session significantly depends on the geographical region of the issuer. That is, Asian peers issue on average less queries than North American and European peers. Furthermore, European peers issue on average more queries per session than North American peers.

Thus, the plot of conditional distributions is an intuitive and easily applicable method for identifying correlations. In fact, this method is in particular useful if a measure represents categorical data instead of quantities, such as the geographical region in the example. Obviously, if the cardinality of $\mathcal{T}$ is large and in particular if Y is a continuous random variable, this method is not applicable. Then, we can either split $\mathcal{T}$ into several ranges, thus, defining bins for the sample space of Y or use the more compact representation of percentile plots explained in the following section.

2.2.2 Percentile Plots

The plot of conditional distributions gets difficult to survey in a single graph when there are many curves representing a conditional distribution, i.e. when the cardinality of $\mathcal{T}$ is large. A more compact representation providing means to identify correlations are *percentile plots* which are used e.g. in [CV01]. Percentile plots do not plot the entire curve of the conditional distributions, but extract significant points only. Theses significant points are specific percentiles, which are commonly chosen as 20%, 50%, and 80%. I.e. a percentile plot shows the value x for which $F_{X|Y}(x|y) = 0.20$, $F_{X|Y}(x|y) = 0.50$, or $F_{X|Y}(x|y) = 0.80$, respectively, over the corresponding values of y. By this method, percentile plots comprise three curves, one for the 20^{th}, one for the 50^{th} and one for the 80^{th} percentile.

Consider the distributions of measure X which takes values in $\mathcal{S} = [0, 1]$ conditioned of measure Y which takes values from $\mathcal{T} = \{2.0, 2.5, 3.0, 3.5\}$ shown in Figure 2.2. In the figure additional lines are included which mark the 50^{th} percentiles for each conditional distribution. The corresponding percentile plot is given in Figure 2.3. It shows that with increasing values for Y the curves for all three percentiles increase. Thus, there is a *positive* correlation between the measures X and Y. The term positive denotes that the values for X increase with increasing values for Y. Similar a correlation is called *negative* when the values for X decrease with increasing values for Y. If the measures were independent, all conditional distributions would be equal and the percentile plot would show three horizontal lines.

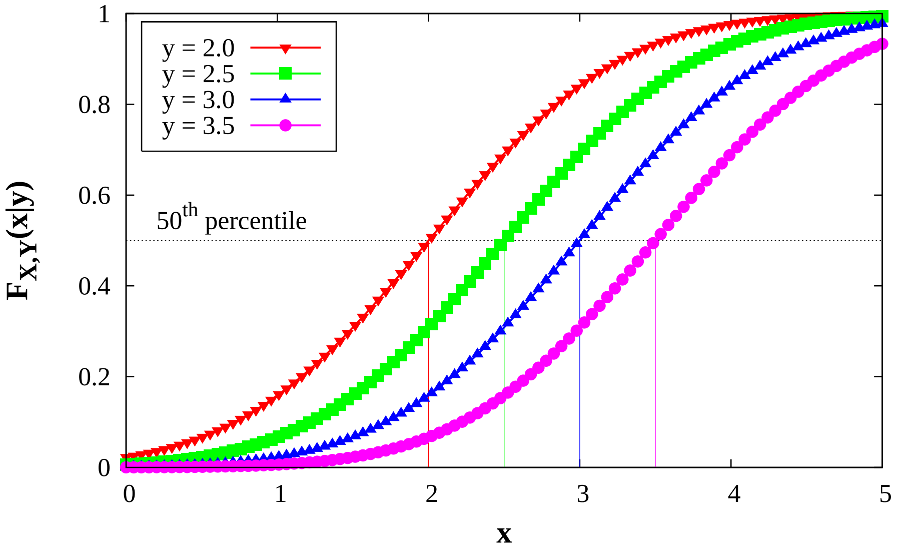


Figure 2.2: Relation between conditional distribution and percentile plot

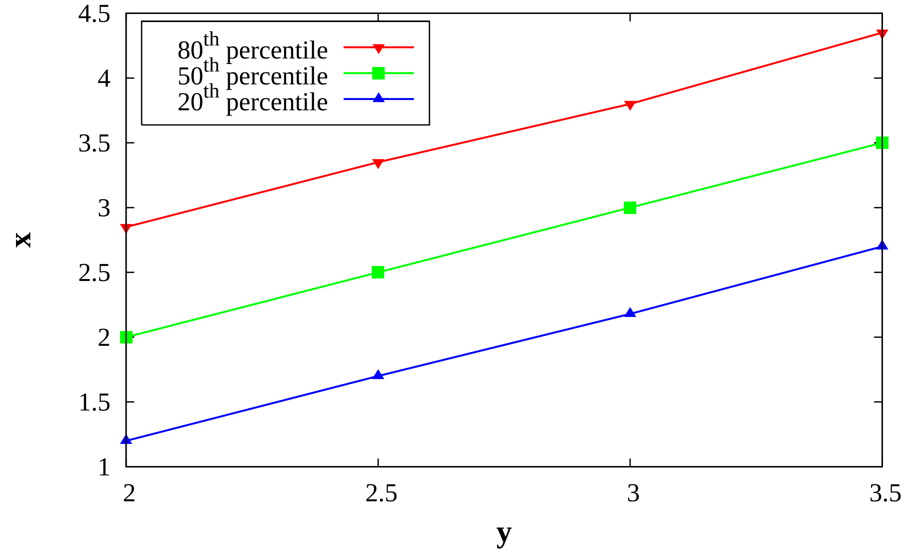


Figure 2.3: Identification of correlations using percentile plot

Since percentile plots aggregate information about each conditional distribution in three data points only, they provide a more compact representation of the correlation structure than the direct plot of the conditional distributions. Furthermore, percentile plots enable the identification of ranges in the domain of Y in which the correlation is significant. In the example of Figure 2.3 measures X and Y correlate linearly. But as will be shown in Chapter 3 correlations may be significant for specific ranges of y values only, whereas other ranges may not show significant correlations. Percentile plots provide means to identify those ranges, by considering the slope of the percentile curves.

2.2.3 Correlation Coefficient

The *correlation coefficient*, also known as *Pearson's correlation* or *product-moment coefficient*, is a quantity to determine how strong two measures correlate. Let x_i and y_i be the values of measure X and Y, respectively, of the i^{th} measured sample, n being the number of measured samples, $\overline{x}$ and $\overline{y}$ being the means of X and Y, respectively. Then the correlation coefficient r [Edw76] is defined as:

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2}} \tag{2.1}$$

The correlation coefficient takes values between -1 and $+1$, where -1 represents perfect negative correlation and $+1$ indicates perfect positive correlation. If the measures are independent, the correlation coefficient calculates to values near zero. Thus, the correlation coefficient captures the correlation between two measures in a single value. Following the "rule of thumb" in [HWJ88], we consider correlations with a correlation coefficient of less than 0.30 as insignificant. Note that the square of the correlation coefficient r^2, denoted as *coefficient of determination* [Edw76], indicates the proportion of shared variance between the measures. That is, a correlation coefficient of 0.50 would yield a coefficient of determination of 0.25, so that 25% of the variation in one measure might be considered associated with the variation in the other measure.

Although the correlation coefficient is attractive because it is the most compact representation of correlations and provides a quantitative measure, it gives no additional insight into the type of correlation. In particular, the correlation coefficient measures linear correlation only. I.e. if two measures are correlated in a non-linear manner (e.g. quadratic), the correlation coefficient would show a zero or low correlation.

2.3 Data Fitting

It is observed in several measurement studies that many processes in nature, social life, and telecommunication systems follow specific recurring patterns which can be described by probability distributions in the forms of mathematical equations. Proper probability distributions reflect the measured behavior accurately and compact and thus, may serve as a stochastic model. Carefully chosen stochastic models are the basis for developing a complete model of the considered system which can be used for getting deeper insight in its functionality and for performance evaluation. Thus, the measurement-based characterization of Internet properties involves finding appropriate stochastic models which accurately reflect the measured data. Since probability distribution functions are in general parameterized, finding a stochastic model for the measured data involves on the one hand the identification of an appropriate probability distribution function (*model selection*) and on the other hand the computation of a corresponding parameter set which best fits to the measured data (*data fitting*). Thus, this section presents a set of probability distributions and an iterative scheme for non-linear curve fitting which are used throughout this thesis.

2.3.1 Stochastic Models for Continuous Distributions

This section recalls the definitions of some well known continuous probability distributions which reflect a wide range of different properties, like heavy-tail behavior, symmetry etc. We focus on continuous distributions here, because most measures observed in the studies of Chapters 3 and 4 are either continuous or take a large number of values, so that they can be well approximated by continuous distributions. Table 2.1 shows the definitions of the *probability density functions* (*pdf*) and the *cumulative distribution functions* (*CDF*) for the distributions used throughput this thesis. A more detailed discussion of the properties of these and further continuous distributions can be found in [JKB94].

The different properties of the considered distributions allow a preselection of distribution functions as candidates for stochastic models. This preselection can

Exponential distribution:	
Probability density function:	$pdf(x) = \lambda e^{-\lambda x}$
Cumulative distribution function:	$CDF(x) = 1 - e^{-\lambda x}$
Parameters:	λ (scale parameter), $\lambda > 0$
Normal distribution:	
Probability density function:	$pdf(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Cumulative distribution function:	$CDF(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt$
Parameters:	μ (location parameter, mean);
	σ (scale parameter, standard deviation), $\sigma > 0$
Lognormal distribution:	
Probability density function:	$pdf(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$
Cumulative distribution function:	$CDF(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \frac{1}{t} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} dt$
Parameters:	μ (scale parameter), $\mu > 0$;
	σ (shape parameter), $\sigma > 0$
Pareto distribution:	
Probability density function:	$pdf(x) = \frac{\alpha k^{\alpha}}{x^{\alpha+1}}$
Cumulative distribution function:	$CDF(x) = 1 - \left(\frac{k}{x}\right)^{\alpha}$
Parameters:	k (location parameter), $k > 0$;
	α (shape parameter), $\alpha > 0$
Weibull distribution:	
Probability density function:	$pdf(x) = \alpha\lambda x^{\alpha-1} e^{-\lambda x^{\alpha}}$
Cumulative distribution function:	$CDF(x) = 1 - e^{-\lambda x^{\alpha}}$
Parameters:	α (shape parameter), $\alpha > 0$;
	λ (scale parameter), $\lambda > 0$

Table 2.1: Definitions of continuous distribution functions

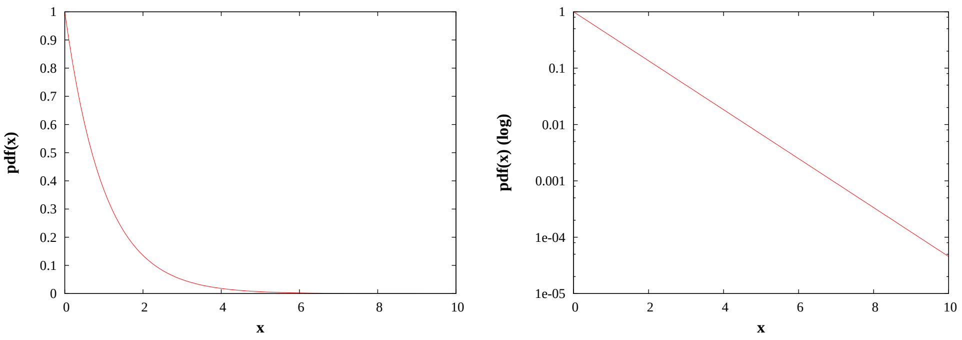


Figure 2.4: Probability density of the exponential distribution with $\lambda = 1$ on linear-linear and log-linear plot

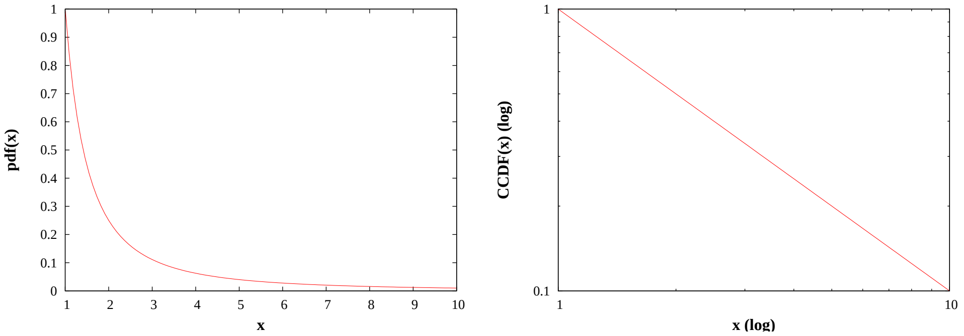


Figure 2.5: Probability density and complementary cumulative distribution of Pareto distribution with $\alpha = 1$ and $k = 1$ on linear-linear and log-log plot

be performed by graphical analysis of the measured data. Consider the distribution of a measure observed in a measurement study. It can easily be tested if it follows an exponential distribution by plotting the measured probability density with linear-scale x-axes and log-scale y-axes. Opposed to the probability density of the exponential distribution on a linear-linear plot shown in Figure 2.4 (left) the log-linear plot shown in Figure 2.4 (right) shows a straight line. Thus, a straight line in the corresponding plot of the experimental data indicates that the measure is distributed exponentially. Similarly, the normal distribution has the characteristic property that its density is symmetric to the mean of the distribution, which is given by the location parameter μ. Thus, a plot of the measured data showing this symmetric behavior indicates that the data may follow a normal distribution.

A further property which can be identified by graphical analysis is the heavy-tail behavior of a distribution. A distribution is said to be *heavy-tailed* [CB97] if $P\{X > x\} \sim x^{-\alpha}$, as $x \to \infty$, $0 < \alpha \leq 2$. Intuitively, this means that a heavy-tailed distribution has a non-negligible probability of showing very large values. The Pareto distribution has this heavy-tail property. Graphically, this property can be identified, if the complementary cumulative distribution function ($CCDF$) of the measured data shows a straight line on the log-log plot for values larger than the location parameter k. Again for illustration purposes Figure 2.5 plots the pdf and the CCDF of the Pareto distribution for $k = 1$ and $\alpha = 1$ on a linear-linear and a log-log plot, respectively.

2.3.2 Least-Squares Regression

A common method used for fitting a stochastic model, i.e. a given probability distribution, to measured data is *least-squares regression*. This approach tries to find a parameter set for a given parameterized function such that the deviation of the measured data points from the curve given by the function is minimal. This deviation, which is meant to be minimized, is given by the sum of the squared vertical deviations (*residuals*) from each data point to the curve. Note that the residuals are squared to avoid cancellations between positive and negative deviations. The sum of the squared residuals is denoted as χ^2. Then least-squares regression for a two-parameter function tries to minimize

$$\chi^2(a,b) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i, a, b)}{\sigma_i} \right)^2$$

where $f(x, a, b)$ denotes the given function with parameters a and b, x_i, y_i are the measured X- and Y-values for data point i with $i = 1, ..., n$ and σ_i is the weight of the i^{th} data point. The weighting parameters σ_i may be used to give more weight to specific data points, e.g. in the case when it is known that there are different measurement errors for different data points. Since weighting requires additional information about the measured data which is usually unknown the weighting parameters may be set to 1 as is done throughout this thesis. Note that least-squares regression is not limited to functions with only two parameters. We concentrate our discussion on this case, because the stochastic models considered here have at most two parameters.

Since both the pdfs and the CDFs of the probability distributions considered throughout this thesis are nonlinear equations, there are in general no closed-form expressions for this minimization problem. Thus, we employ the most widely used iterative scheme for least-squares regression, namely the Levenberg-Marquardt algorithm [BW88]. This algorithm iteratively calculates $\chi^2(a, b)$ and changes the parameter values a and b to decrease $\chi^2(a, b)$. The iterations are stopped when either a specified number of iterations has been processed or the parameters converge to a final parameter set where χ^2 changes by a factor less than a specified epsilon from one iteration to the next. The value of epsilon thus specifies a certain accuracy for finding a parameter set with minimum χ^2. Throughout this thesis, the Levenberg-Marquardt algorithm implemented in the plotting tool `gnuplot`

[WK05] is used for calculating the set of parameters for which a given probability distribution fits best to the measured data.

In general, the most applicable probability distribution for an observed measure is not known in advance. Therefore, we need to identify the distribution which best represents the measured data. Often this cannot be done solely by the help of the considerations about the specific properties of the different distribution functions stated above. Then least-squares regression is performed for all potential probability distributions and the best suited model is found by means of the *root-mean-square of residuals (rms)* [KK62] denoted as Δ. The root-mean-squares of residuals is defined as

$$\Delta = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - f(x_i, a, b)\right)^2}{n}}$$

As the name suggests, Δ is the rooted mean of the squares of residuals. I.e., Δ approximates the average deviation of the measured data points from the curve. This quantity provides a simple but effective measure for the goodness-of-fit. The smaller Δ the better the model fits to the measured data. Thus, in cases where multiple probability distributions are candidates for matching to a measure, the distribution with the smallest Δ is chosen.

Chapter 3

Characterization of the Query Behavior in P2P File Sharing Systems

In the last few years Peer-to-Peer (P2P) file sharing has become the second killer application in the Internet besides World-Wide-Web (WWW) browsing. With the increasing popularity of such systems, which enable the sharing and exchange of files between Internet hosts, the need for novel algorithms and protocols to improve scalability and efficiency becomes crucial. To evaluate the performance of these approaches, workload models are required which accurately represent current usage of these systems. Previous work on workload characterization of P2P file sharing systems has not analyzed how individual users search for files. Thus, this chapter presents a synthetic workload model for the query behavior which is based on a comprehensive measurement study in the Gnutella overlay network. The measurement results reveal the importance of filtering out automated queries, which are not explicitly triggered by the user, for extracting user behavior. Considering only user generated queries, a detailed characterization of the measured data builds the basis for developing a synthetic workload model which accurately reflects the heterogenous query behavior of individual peers. The main results presented in this chapter have been published in the proceedings of the *Internet Measurement Conference (IMC 2004)* [KLVW04] and with a different focus in the proceedings of the *GI/ITG Conference on Measuring, Modeling and Evaluation of Computer and Communication Systems (MMB 2004)* [KLW04b].

3.1 The Gnutella P2P File Sharing System

Gnutella [Gnu04] is a protocol specifying a fully distributed system for sharing files between hosts in the Internet. Since the Gnutella protocol is an open source project, the Gnutella system constitutes numerous client implementations and is currently one of the most popular P2P file sharing systems. Gnutella peers, often called *servents* (as a mixture of *serv*er and cli*ent*), construct an overlay network, i.e. a network of application layer connections, for enabling communication between peers. This overlay network (denoted as *Gnutella network*) is used to route signaling traffic, i.e. messages for maintaining overlay membership and messages for searching files, from one peer to another.

For its operation, the Gnutella protocol specifies four message types. Messages of types `PING` and `PONG` are used to maintain overlay membership. `PING` messages are used to actively probe the network, i.e. to ask for information about other active peers. Peers receiving a `PING` message respond with one or more `PONG` messages containing the IP address and port number of active peers and the amount of data they share. Messages of type `QUERY` contain a set of keywords in the name of files a user is searching for. If a peer receiving a `QUERY` message shares one or more files that match to the query string, i.e. the set of keywords in the query string is a subset of the keywords in the name of one or several shared files, it responds with the fourth message type, `QUERYHIT`. This response message specifies the IP address and port number of the responding peer as well the names and sizes of the matching files.

The operation of the Gnutella system is as follows: For connecting to the Gnutella network at application startup, peers have to connect to an initial overlay member. The determination of the IP address and port number of such a peer (*bootstrapping*) is performed by contacting one of several well-known hosts which are always online and maintain a list of active peers. Furthermore, on application exit the peer stores known peers in a disk cache which can be used for bootstrapping at the next application startup. Due to the dynamics of the peers participating in the Gnutella network, there are frequent leaves and joins of peers. Thus, to preserve network integrity, each peer maintains overlay connections to a number of other peers. This number of overlay connections is kept between two parameters `min_conn` and `max_conn` which can be adjusted by the user. If the number of overlay connections is below `min_conn`, the peer establishes additional connections. Candidates for further overlay connections are obtained by sending `PING` messages to already connected (*neighboring*) peers and inspecting `PONG` responses. Connection requests by other peers are accepted until `max_conn` connections are active.

For searching files in the Gnutella network, peers generate a `QUERY` message which is sent to each of its directly connected peers (*one-hop peers* or *neighbors*) in the overlay network. These neighbors forward the message to their neighbors, thus flooding the network. Forwarding a `QUERY` message more than once is prevented by storing the query's global unique identifier (GUID) in a routing table, along with the identity of the neighboring peer that the query is initially received from. The maximum number of overlay hops that a `QUERY` message may transit is specified by a time-to-live (TTL) field, which is set when the message is generated. Typically the TTL is set to an initial value of seven. The field is decremented each time the message is forwarded, and the message is not forwarded if TTL is equal to zero. Thus, the TTL is similar to the corresponding field used e.g. in the `ping` program. To determine how far a message has traveled through the overlay network, a hop count field with an initial value of zero is incremented before forwarding. Peers receiving the `QUERY` message respond with a `QUERYHIT` message if they share matching files. This message is transferred to the inquiring peer on the reverse overlay path that the query message was routed to the responding peer. This is done by using the routing table that specifies the next hop in the reverse path for each GUID. According to the protocol specification, a GUID is deleted from the routing table after a specified time, typically after 10 minutes.

In its initial version as described above, the Gnutella protocol specified a completely decentralized P2P system in which all peers were considered equal. However, for combining the advantages of centralized and decentralized P2P systems the current Gnutella protocol version v0.6 has a *hybrid* structure, similar to the proprietary FastTrack protocol used e.g. by KaZaA [Sha04]. Hybrid systems distinguish between powerful peers (in the sense of CPU power and bandwidth of the Internet connection), so called *ultrapeers* or *super nodes*, and less powerful peers denoted as *leaf nodes*. Note that we will use the denotations *peer* and *node* interchangeable throughput this chapter. In hybrid systems an ultrapeer acts as indexing server for a number of leaf nodes and shields these less powerful peers from the vast amount of search queries. Thus, clients implementing the current version of the Gnutella protocol decide if they act as ultrapeer or leaf node. In ultrapeer mode they maintain muliple simultaneous overlay connections to other ultrapeers and accept connection request from leaf nodes. Leaf nodes connect to a small set of ultrapeers and only receive `QUERY` messages which likely match to locally stored files.

When a peer locates files it wishes to download using the mechanism described above, the actual file transfer is performed by a direct TCP connection between the inquiring and the responding peer (peer-to-peer) using the well-known hypertext transfer protocol (HTTP). Thus, the overlay network is solely used for locating available files shared by the peers. Note that the Gnutella protocol is extensible, so that software clients implementing the Gnutella protocol may provide additional features not described here.

3.2 Previous Results on Characterization of P2P File Sharing Systems

The full characterization of P2P file sharing systems incorporates four related aspects: *(i)* The characterization of the overlay topology includes how the overlay network is constructed and how it evolves. *(ii)* The characterization of available files includes how many files are shared by the peers together with their size and replication, i.e. how often a specific file is shared by different peers. *(iii)* The characterization of the download behavior includes which files are downloaded from which source peer and how the download evolves. *(iv)* The characterization of the query workload includes which queries the peers issue at what time for locating available files.

Characterizations of Overlay Topology

Several papers report measurement studies of the overlay topology of P2P file sharing systems. Ripeanu, Iamnitchi, and Foster [RIF02] measured the Gnutella network using a crawler between November 2000 and May 2001. They discovered that the connectivity distribution, i.e. the number of overlay connections to other peers, significantly changed during that time period where the network grew about 25 times. Furthermore, they observed that the overlay topology substantially differs from the topology of the underlying physical network, resulting in inefficient network usage.

Bhagwan, Savage, and Voelker [BSV03] analyzed the availability of peers in P2P systems by measurements in the Overnet P2P file sharing system. They found that the availability of peers depends on the time of day and that there are short-term daily joins and leaves of individual peers as well as long-term arrivals and departures of peers.

Saroiu, Gummadi, and Gribble [SGG02] measured the Napster and Gnutella file sharing systems in order to characterize the peers in terms of network topology, measured bottleneck bandwidth, network latency as a function of bandwidth, session duration, number of shared files, size of shared files as a function of the number of shared files, and number of downloads as a function of peer bandwidth. They identified different classes of peers and propose that different tasks in a P2P system should be delegated to different peers depending on their capabilities.

A detailed characterization of the modern Gnutella overlay topology is provided by Stutzbach, Rejaie, and Sen in [SRS05]. By capturing more than 18,000 snapshots of the Gnutella network between April 2004 and February 2005 the authors analyzed graph-related properties of individual snapshots as well as the dynamics of the overlay. The analysis revealed that the power-law distribution of peer connectivity reported in previous work is caused by measurement artifacts. Furthermore, the Gnutella protocol forms an "onion-like" connectivity among peers with a stable and densely connected core-overlay consisting of long-lived peers.

Characterizations of Shared Files

The characterization of files shared in P2P file sharing systems is the goal of further publications. Adar and Huberman [AH00] measured the Gnutella system and found a significant fraction of free rider sessions, which download files from other peers but don't share any files. They argue that free riding degrades the system performance and, therefore, propose to incorporate mechanism to minimize free riding in future file sharing systems.

An analysis of locality in shared files and downloads is provided by Chu, Labonte, and Levine [CLL02]. They periodically collected shared file lists from Napster and Gnutella peers over a period of several weeks. The analysis of this data showed that both file locality as well as download locality fit to a log-quadratic distribution.

Sripanidkulchai, Maggs, and Zhang [SMZ03] discovered interest-based locality in the files downloaded by Web, KaZaA, and Gnutella clients by examining five different one-day traces. A significant fraction of users tend to share similar sets of files. They showed that this locality can be exploited to improve the search performance in P2P systems by inventing shortcuts in the overlay topology based on common interests.

This interest-based locality is also observed by Fessant et al. [FHKM04] in the eDonkey P2P file sharing system. Furthermore, the authors showed by measurements taken during the first week of November 2003 that there is also a geographical clustering of shared files, i.e. specific files are replicated more often in a specific geographical region than in other regions.

Conducting various measurements in the Gnutella network in June 2005 Zhao, Stutzbach, and Rejaie [ZSR06] discovered the characteristics of the shared files in three ways: static analysis, topological/geographical analysis and dynamic analysis. They compare the findings of the recent measurements with those of the previous work [AH00], [CLL02], [SMZ03], [FHKM04] and additionally show that file preferences are correlated with the geographical region but independent of topological properties. Furthermore they show that popular files experience further variation in their popularity than unpopular files.

Characterizations of File Downloads

The file download process in P2P file sharing systems is analyzed in a number of papers by means of capturing packet traces. Saroiu et al. [SGD$^+$02] measured the HTTP headers of Web, Akamai and P2P traffic in order to compare the characteristics of these three content delivery systems. Note that HTTP is used by KaZaA as well as by Gnutella file transfers. They found significant differences in the file size distributions, the bandwidth consumed by the file transfers, and the file types among the three systems.

In [GDS$^+$03] Gummadi et al. collected and analyzed a 200-day trace of KaZaA traffic. They characterized active session length, size of downloads, and evolution of object popularity. Similarly, Sen and Wang [SW02] characterized traffic volume, distribution of time between downloads, and active session length based on a three month flow-level measurement of FastTrack, Gnutella and DirectConnect traffic. By collecting data in the OpenNapster system (an open source successor of the Napster system) Ng et al. [NCR$^+$03] analyzed the peer selection technique for downloading files. They showed that simple enhancements based on basic measurement techniques like RTT probing and bottleneck bandwidth probing significantly improves the download performance.

Characterizations of Query Workload

There is only little research done in the field of characterizing the query behavior of peers in a P2P file sharing system. Although the search process used in P2P file sharing systems is subject of various research projects e.g. [CRB$^+$03], [SMZ03], [LCC$^+$02] there are only few measurements of the query behavior. Note that object popularity as analyzed e.g. in [GDS$^+$03] is related but different to query popularity. That is, while query popularity denotes the popularity of search messages, object popularity denotes the popularity of download requests. The main difference is, that queries may match to several different files and are received by multiple peers. In contrast, download requests correspond to specific files and are received only by peers sharing this file. Thus, query messages do not necessarily correspond to download requests. A peer may issue multiple download requests after receiving responses to a query message. Or it may issue no download request if the results are not satisfactory.

Sripanidkulchai [Sri01] analyzed the popularity of queries in the Gnutella network. She showed that the popularity of Gnutella queries follows a Zipf-like distribution, where the relative probability of the i^{th} most popular query is proportional to $1/i^{\alpha}$, for values of α between 0.63 and 1.24. Sripanidkulchai proofs that caching of query results can reduce the network traffic up to a factor of 3.7. Similarly, Krishnamurthy, Wang and Xie [KWX01] observed a Zipf-like distribution of query popularity in their Gnutella traffic trace.

Both, [Sri01] and [KWX01] do not aim at providing a detailed workload model and only consider the aggregate query popularity. Thus, their work cannot be used for a synthetic workload model of individual peers. In contrast, this chapter provides a detailed characterization of the query behavior of individual peers. Furthermore, a synthetic workload model comprising appropriate distributions and parameters is directly applicable for the evaluation of existing and novel P2P system designs. In contrast to previous work in this area, we consider all aspects necessary for generating representative synthetic workload. This includes incorporating time of day effects, correlations to the geographical region, distribution of interarrival times between subsequent queries of the same peer, just to name a few. Furthermore, we separate system-independent peer behavior from system-dependent behavior in order to develop a query workload model which is widely independent of the measured system.

3.3 Measurement Methodology for the Gnutella Overlay Network

3.3.1 Measurement Setup

For the analysis of peer behavior in P2P file sharing systems, we set up a client node in the Gnutella overlay network [Gnu04]. Since the Gnutella protocol specification is publicly available, there are a number of client implementations. To perform the measurements in the Gnutella network, we modify the open-source Gnutella client `Mutella` [Mut04], to record a trace of all Gnutella messages routed through the node. We denote this node as *measurement peer*. For each message the trace includes a line of text containing the time when the message was received, the type of the message, and for each message type a well formatted textual representation of the data contained in the message such as query string, GUID etc. Furthermore, the trace contains additional entries for the establishment and termination of TCP overlay connections. These entries include, amongst others, the timestamp of the event as well as the IP address, port number and client version of the connecting or disconnecting peer. Before writing the recorded data stream to the hard disk of the measurement host, the stream is compressed in order to reduce disk space requirements. We conduct only passive measurements; that is, we do not generate messages actively, in order to minimize the disturbance of the current network traffic by the measurement. Note that the clock of the measurement host is synchronized to stratum 1 time servers using the network time protocol (NTP) [Mil92]. Since the unit of the timestamps used in the traces is 1 second, which is sufficient to characterize the query behavior of P2P file sharing users, the measured data is accurate with respect to the time related measures (see Section 2.1).

To obtain a reasonable sampling of the network traffic in the traces, we specify that the measurement client will run in ultrapeer mode and maintain up to 200 connections to other peers simultaneously. This results in more than four million

Measure	Value
Trace period	3/15/04 - 4/23/04
Number of QUERY messages	34,425,154
Number of QUERYHIT messages	1,339,540
Number of PING messages	27,159,805
Number of PONG messages	17,807,992
Number of direct connections	4,361,965
QUERY messages with hop count = 1	1,735,538

Table 3.1: Summary of trace characteristics

measured direct peer connections during the forty day measurement period, as shown in Table 3.1. Approximately 40% of the connections are from peers that are running in ultrapeer mode, and 60% are from leaf peers or peers that do not support the ultrapeer mechanism (*normal peers*). Thus, all types of nodes are well represented in the measured workload. Overall, the measured Gnutella signaling traffic constitutes more than 10 GBytes of compressed trace data.

For convenience, the measurement node is located at the University of Dortmund. However, as will be shown in Section 3.3.4 below, the directly connected peers are scattered around the globe with proportions of one-hop peers in North America, Europe, and Asia that are approximately the same as the corresponding proportions of the total peer population in each of the three continents. Since the construction algorithm of the Gnutella overlay network does not contain any geographical bias in the peers that are directly connected, we hypothesize that the placement of the measurement node does not impact the measured behavior of the peers. Section 3.3.4 provides quantitative measures that confirm this hypothesis. Furthermore, we make sure that the measurement is not biased by filtering rules and traffic shapers located at the Internet gateway of the campus network. This is done by assigning a special IP address to the measurement node which is not affected by the traffic control mechanisms of the Center for Communication and Informationprocessing (Hochschulrechenzentrum).

3.3.2 Measuring Peer Characteristics

This section shortly outlines how the most important characteristics of the measured data are measured. Since this chapter deals with the query behavior of individual peers, the queries observed in the measurement have to be related to their specific issuer. Unfortunately, a QUERY message does not include the IP address or any other tag that can be used to identify the peer that generated the query. However, each QUERY message generated by a user of a Gnutella client that is directly connected to the measurement peer has a hop count equal to one, and the IP addresses of these directly connected peers are known from the TCP connections in the overlay. Since each QUERY message that is generated at a client is sent to each directly connected peer (at least to all non-leaf peers), the measurement node will receive every QUERY message from a directly connected peer. We can thus relate QUERY messages to the corresponding issuer for each connected peer session that has a distance of one hop. As stated in Table 3.1, the trace comprises more than 1.7 millions QUERY messages, for which the issuer can be uniquely identified. The remainder of this chapter is based on these QUERY messages. To determine the geographical region of a peer we use the GeoIP database [Max04] for mapping IP addresses to continents.

A further important measure is the peer's session duration. There are no Gnutella messages to indicate the start of a new client session. However, a connected session starts when the Gnutella handshake between the measurement peer and the one-hop peer is completed. Since the measurement client session runs without interruption until the end of the measurement period, the termination of the TCP connection to a one-hop peer indicates the end of the one-hop peer's session. We note that many Gnutella clients do not terminate an overlay connection by sending a BYE message according to the Gnutella specification. Instead, most clients simply stop sending messages over the connection or cancel the TCP connection. When the measurement peer detects that a connection is idle for 15 seconds it sends a single PING message to the one-hop peer. If no response is received after another 15 seconds the measurement peer will close the connection. Thus, we will overestimate the end of most connected session durations by approximately 30 seconds.

According to the Gnutella protocol, a query string matches to a shared file if the set of keywords extracted from the query string is a subset of the keywords extracted from the filename. That is, Gnutella clients split query strings and filenames at delimiter characters and compare the resulting sets of keywords. We use this definition when measuring the number of distinct queries observed at the measurement peer. Queries are assumed to be identical if they contain the same set of keywords. We have also verified by inspection that the great majority of the top 100 queries are each for different files, rather than being variations of keywords for the same files.

3.3.3 Filtering Gnutella System Behavior

By analyzing outliers in the trace file according to Section 2.1, we discovered several anomalies in the queries received from some one-hop peers. Examining the content of the User-Agent header exchanged during the handshake at connection establishment, we determined that certain types of anomalies could be attributed to peers running a specific client implementation. Since our objective is to characterize the user workload rather than the behavior of the P2P system software, we discard the following types of query messages that are automatically issued by particular Gnutella client implementations to improve system responsiveness:

1. *QUERY message with the SHA1 extension.* The client software uses the SHA1 hash sum to identify a specific file that is already known. Thus, this query does not indicate the user's interest in a new file, but rather a search for additionally sources to continue a file download.

2. *QUERY message with a query string that has already been observed within a client session.* Many Gnutella clients provide features for automatically resending a query in order to improve search results. These repeated queries indicate that the system is searching for further results, rather than user behavior.

3. *QUERY message from a session that is connected for less than 64 seconds.* A large fraction of one-hop peers (i.e. 29%) disconnected in less than 10 seconds and another significant fraction (32%) disconnect during the next 20-25 seconds. A total of about 70% of connections terminate in less than 64 seconds. Such frequently occurring quick disconnects are likely due to system software decisions to disconnect from the measurement peer (for unknown reason) rather than user behavior. They could partly be caused

by Gnutella crawlers, which try to generate a map of the Gnutella topology [SRS05]. However, these connections do not provide meaningful data for characterizing user behavior, thus they have to be discarded. Since other specific connection durations are not observed with unusual frequency, sessions longer than 64 seconds are assumed to end due to user session termination.

Filtering sessions with a length less than 64 seconds will eliminate anomalies in statistics for session duration and number of queries issued per session. In addition, the following query messages are sent by some peers soon after connecting to the measurement peer:

4. *QUERY messages with interarrival time of less than 1 second*, and
5. *QUERY messages with identical interarrival times.* We found some peers that issued query messages in regular intervals, e.g., 10 seconds.

Each of these queries indicates with high probability automated client behavior. They appear to be automated re-queries for queries that were issued by the user prior to connecting to the measurement peer. The client regularly re-submits QUERY messages for each search task which the user does not close. I.e. many clients allow to simultaneously open multiple search tasks in tabs. In the case of query interarrival times of less than 1 second, the client automatically re-queries all open search tasks in one bulk. Other clients automatically re-query open search tasks in a round-robin fashion, resulting in QUERY messages with identical interarrival times. Although queries identified by rules 4 and 5 were generated automatically by the system software, the user query that was issued before the client connected to the measurement peer is important. Thus, we include these queries in the measures of query popularity distribution and number of queries per session, but not in the measure of query interarrival time since the observed arrival time was determined by the system software.

Table 3.2 shows the number of queries that are discarded when each of the first three rules is applied in sequence, and the number of queries that are not counted in the measure of query interarrival time due to rules four and five. We note that the number of queries discarded by each of the first three rules is substantial. For example, nearly half the queries are discarded by the second rule, which identifies queries that are repeated by the system to obtain further results, rather

	Rule	**# Queries**	**# Sessions**
Number of sessions and query messages from 1-hop neighbors		1,735,538	4,361,965
1	Ignore query messages with empty keywords and SHA1 extension	410,513	
2	Ignore query messages with identical query string issued by the same peer within a session	841,656	
3	Discard sessions with session length of less than 64 seconds	310,164	3,053,375
Final number of QUERY messages and sessions considered		173,195	1,308,590
4	Ignore query messages from a specific peer with query interarrival time of less than 1 seconds	77,058	
5	Ignore subsequent query messages from a specific peer with identical interarrival times	14,715	
Final number of QUERY messages considered in query interarrival time measure		81,432	

Table 3.2: Number of queries matched by filter rules

than queries that are part of the user workload. Considering the large fraction of automatically generated queries, we conclude that it is essential to apply the filter rules in order to characterize the system-independent query behavior of users. Nevertheless, there are still a substantial number of queries and connected sessions that are analyzed to obtain the user workload characterization.

3.3.4 Representativeness of One-hop Peers

Since we can only measure the query behavior of peers that are directly connected to the measurement node, we examine two measures that are consistent with the hypothesis that the large number of one-hop peer sessions are representative of all peer sessions in the system.

The first measure is the geographical distribution of the one-hop peers as compared to all peers. To measure the geographical distribution of all peers, we determine the distribution of the IP addresses in all PONG and QUERYHIT messages that are recorded at the measurement node. To measure the geographical distribution of the one-hop peers, we determine the distribution of the IP addresses for all connected sessions. Figure 3.1 provides the fraction of one-hop peers and the fraction of all peers, in each of the three geographical regions where

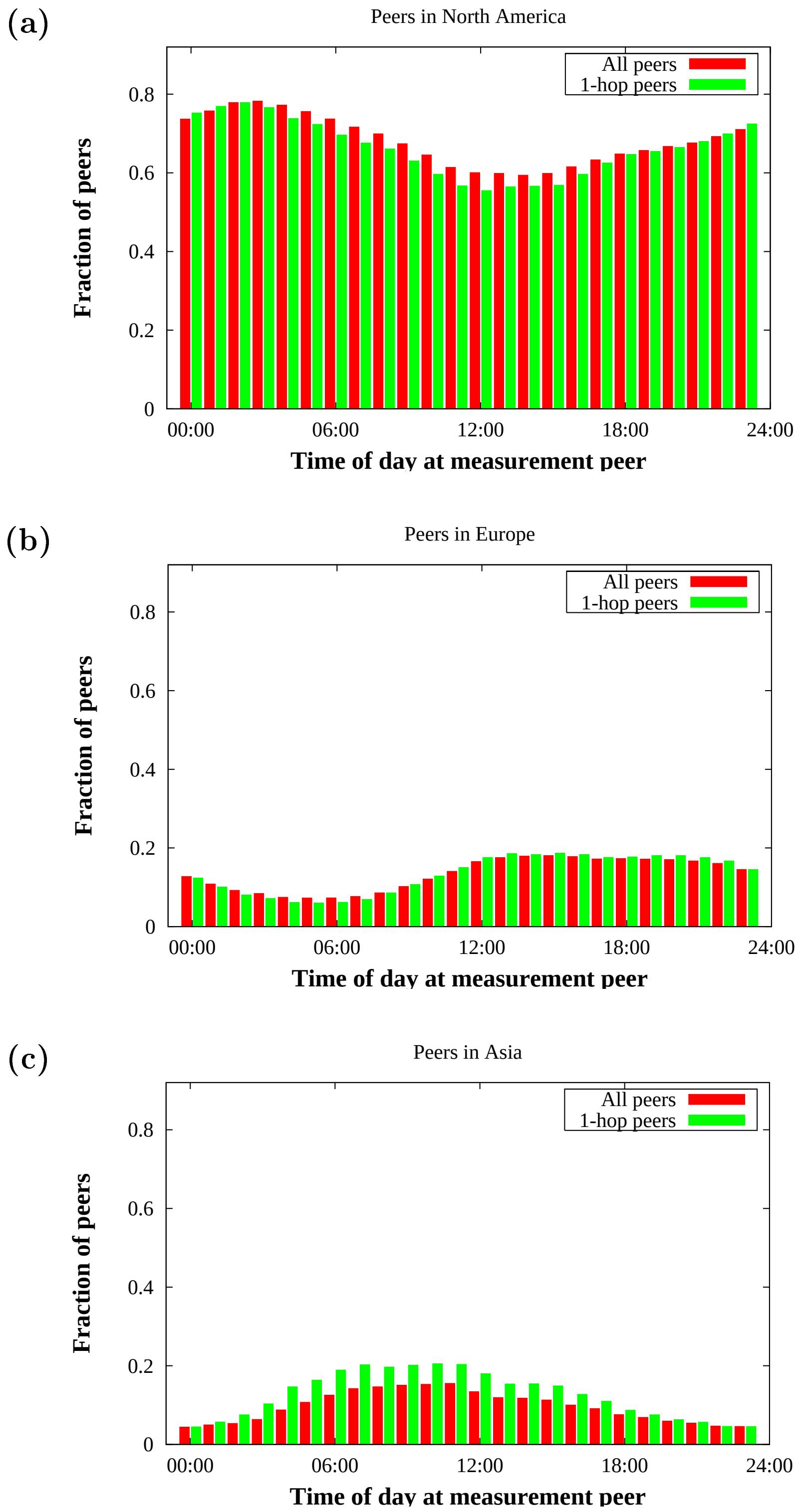


Figure 3.1: Representativeness of 1-hop peers: geographical distribution

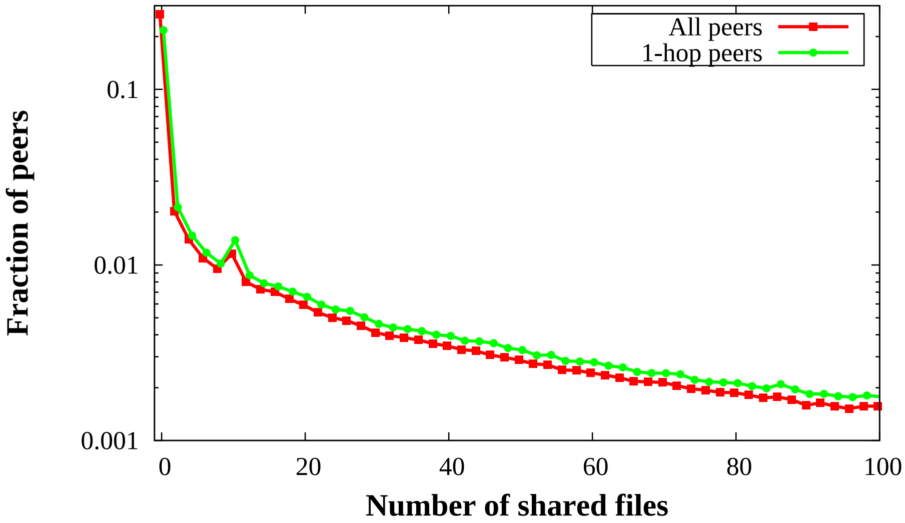


Figure 3.2: Representativeness of 1-hop peers: number of shared files

most peers are located (North America, Europe, and Asia) during each one-hour interval of a day. The value for each one-hour bin is an average over the entire trace. We observe that the geographical distribution of one-hop peers is nearly the same as the geographical distribution of all peers, although there is a slightly higher fraction of one-hop peers in Asia and a slightly lower fraction in North America during the daytime hours at the measurement node. As we will show in Section 3.4, Asian peers tend to maintain shorter sessions than the peers in the other two continents, so the number of Asian peers that are more distant than one hop from the measurement node may be somewhat underestimated. We, thus, conclude that the Gnutella client software connects to one-hop peers that are widely distributed and appear to be randomly selected with respect to geographical region.

In a second experiment, we analyze the number of shared files as reported in PONG messages from all peers and in PONG messages from one-hop peers. Figure 3.2 plots the fraction of each class of peers that report number of shared files from zero to one hundred. We observe that one-hop peers are reasonably representative of the total peer population with respect to the number of shared files. In the next section, we characterize the behavior of the one-hop peers, noting that the measures presented in Figures 3.1 and 3.2 are consistent with the hypothesis that the one-hop peers are representative of the total peer population.

3.4 Characterization of Query Behavior

Each connected (one-hop) peer session can be classified as either *active* or *passive*. Active peers send at least one query in order to locate files for download. Passive peers are connected in the overlay network but issue no queries. These peers constitute an important component of a realistic workload because they don't generate any query load and because they form part of the overlay network that forwards and responds to queries. Thus, our characterization includes both types of peers.

A key goal is to characterize all distributions needed for generating a synthetic workload that accurately captures the query behavior of individual peers. Key correlations among the workload characteristics need to be represented in the synthetic workload. We, thus, begin in Section 3.4.1 by characterizing the fraction of peers from each of the three continents where most peers reside, as a function of the time of day. Section 3.4.2 characterizes the total query load from each of the three continents as a function of the time of day, and identifies peak periods for each continent. The session characteristics that need to be represented in a synthetic workload will then be conditioned on geographical location and/or on high-load periods of the day, for whichever characteristics are found to be heterogeneous in either of those domains.

To characterize connected peer sessions, we analyze the fraction of peers that are passive (Section 3.4.3), the distribution of the session duration for passive peers (Section 3.4.4), the distributions of the number of queries per session and the query interarrival times for active peers (Section 3.4.5), and the query popularity distribution (Section 3.4.6). The correlations among these session characteristics and the geographical location of the peer or the time of day are determined as each characteristic is analyzed. Significant correlations are captured in the form of conditional distributions, so that the correlations can easily be represented in a synthetic workload. Model distributions for the important session characteristics are also given, along with representative graphs that illustrate the close fit between the model distributions and the measured distributions. The algorithm for generating a synthetic workload from the measured characteristics is summarized in Section 3.4.7.

3.4.1 Geographical Distribution

As first measure of the workload characterization, we analyze the geographical distribution of peers conditioned on time of day at the measurement node. Curves for the average fraction of peers from each continent during each hour of the 24-hour day have been presented in Figure 3.1. The fraction of peers from each continent on the outlier days (i.e. days with largest or smallest fraction from each region) only differ by about 5% in absolute value from the averages shown in the figure. Thus, the fractions of peers from each region during each hour in Figure 3.1 are approximately representative of the relative mix of peers during each hour on any given day.

We observe from Figure 3.1 that the relative fraction of peers from each geographical region changes modestly as a function of time of day. For example, the fraction of North American peers decreases from about 80% to about 60% during the hours of 3:00 to 12:00 at the measurement node, and then rises gradually back to 80% between 12:00 and 3:00. Note that the x-axis in Figure 3.1 shows the local time at the measurement peer, i.e. CET. European and Asian peers constitute much lower fractions of the Gnutella peers. The largest fraction of European peers, close to 20%, is observed during noon to midnight. At about 6:00, their fraction constitute only about 6%. Similarly, the highest fraction of Asian peers (about 13%) occurs during the afternoon and evening hours in Asia, i.e. between 7:00 and 11:00 at the measurement peer. During the early morning hours only about 4% of the peers are from Asia. Peers from other geographical regions or with unknown origin constitute approximately 5-10% of the peers. To create a synthetic workload, the interesting mixes of peers from North America, Europe, and Asia (respectively) are for instance: 75%, 15%, 5% at 00:00, or 80%, 5%, 5% at 3:00, or 60%, 20%, 15% at 12:00. In the remainder of this chapter, we characterize the peers in the three continents where most peers reside.

3.4.2 Periods of Peak Load

To analyze potential correlations between the various workload measures and the time of day, it is useful to identify periods of time that have high and low query activity for each geographical region. To accomplish this, Figure 3.3 plots the number of queries received from the one-hop peers from each geographical region in bins of 30 minutes as a function of time of day. The values of each bin are

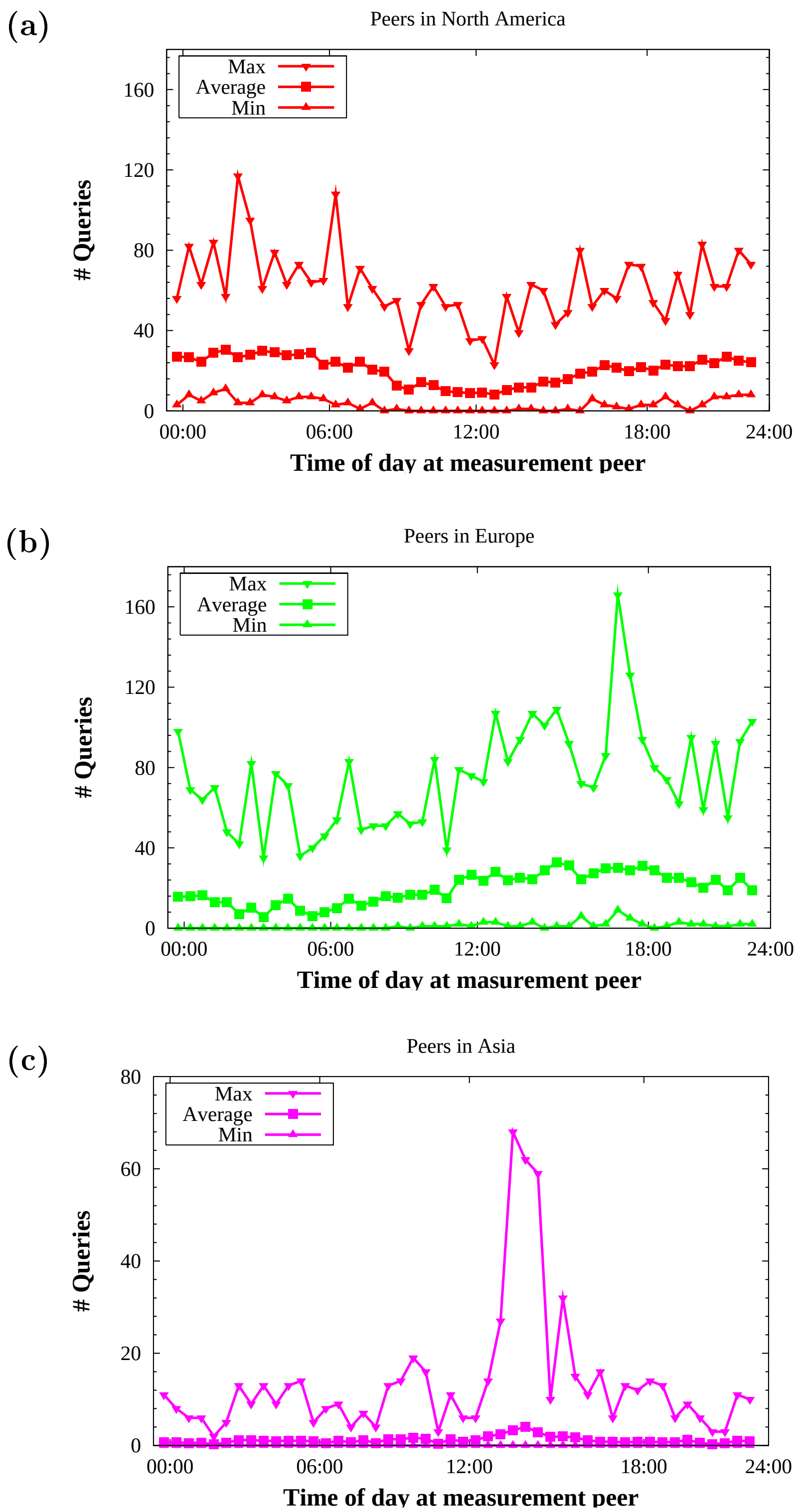


Figure 3.3: Load measured in number of queries vs. time (30 minutes bins)

averaged over the entire measurement period. Except for the Asian peers, the average curves for each region show a similar correlation to time of day as Figure 3.1. In particular, we identify the following key periods from Figure 3.3:

03:00-04:00 peak in North America, sink for Europe

11:00-12:00 sink for North America, peak for Europe

13:00-14:00 sink for North America, peak for Europe, peak for Asia

19:00-20:00 joint peak for North America and Europe.

The minimum and maximum curves indicate a high variance for each bin. Plots of the number of queries during each hour of an outlier day - i.e. a day with the minimum or maximum total number of queries from the geographical region - show that the number of queries is not consistently low or high during each 30-minute interval, but instead varies greatly from one interval to the next. This is due to statistical fluctuations in a relatively small sample size during each interval. That is, statistically, some peer sessions issue larger or smaller number of queries than the average, causing the total number of queries during the interval to differ significantly from the average.

3.4.3 Fraction of Passive Peers

We plot the fraction of passive peers versus time of day in Figure 3.4. In this figure, we count the number of peer sessions that start in a 1-hour interval that issue no queries (during the entire session) and calculate the ratio to all sessions that start in the same hour. The average for each 1-hour interval is computed over the entire measurement period. We observe that the fraction is almost the same for each geographical region, with about 80% to 85% for North America, 75% to 80% for Europe, and 80% to 90% for Asia. Furthermore, the fraction of passive peers fluctuates only by about 5% over time of day. Comparing similar graphs with averages for each bin calculated over the first and the second half of the measurement period, the fraction of passive peers does not change. We conclude from these results that the fraction of passive peers is approximately independent of time of day and of multiple-day periods. Recall that queries which are not issued by users but by the client software are not counted in this measure. Thus, passive peers as defined here may well issue automated re-queries, so that the

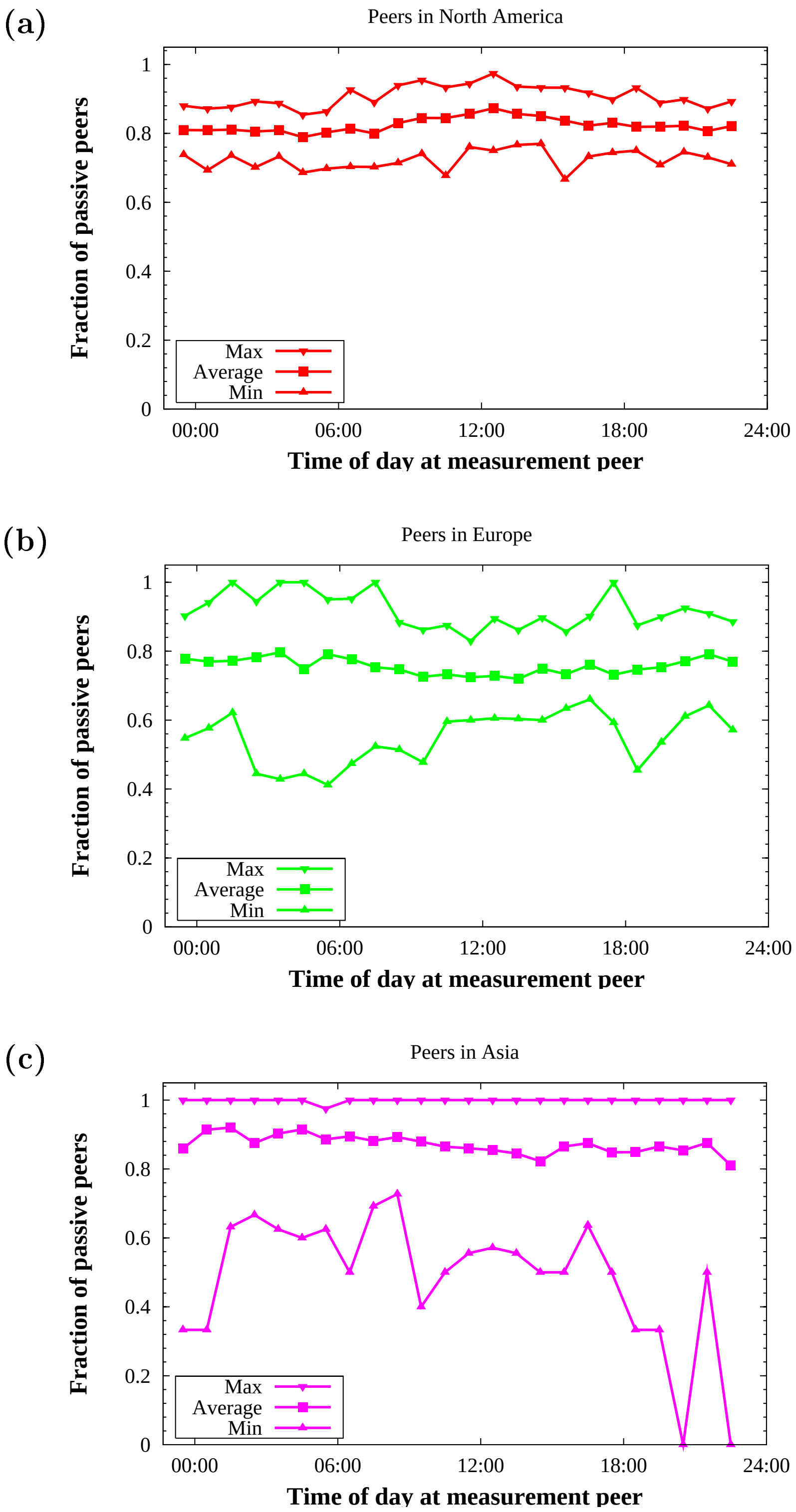


Figure 3.4: Fraction of connected peers that are passive

motivation to start the client and connect to the Gnutella network for those peers likely is to resume unfinished downloads.

3.4.4 Connected Session Duration for Passive Peers

As a passive peer does not send queries, the connected session duration is given by the time during which the peer maintains at least one connection to another peer. To check the correlation between session duration and geographical region, Figure 3.5 (a) plots the complementary CDF (CCDF) of session duration broken down to geographical region. We observe that session duration shows a significant correlation to geographical region. For instance, in Asia 85% of the sessions are shorter than 2 minutes, in North America and in Europe only 77% and 60% are shorter than 2 minutes, respectively. Sessions of an intermediate duration between 2 and 200 minutes constitute 12% in Asia, 19% in North America, and 34% in Europe. Longer sessions make up 3% in Asia, 4% in North America, and 6% in Europe. Note that session durations between 17 and 50 hours account for 1% of the sessions in each geographical region, indicating that a considerable fraction of the peers stays online for a very long time without generating queries. Considering the impact of multiple-day periods, Figures 3.5 (b) and 3.5 (c) show that the distribution of session duration is nearly identical in the first and the second half of the measurement period.

To analyze correlations between session duration and time of day, Figures 3.6 (a), 3.6 (b) and 3.6 (c) plot the CCDF of session duration for sessions starting in each of the important periods. We observe that for European peers session duration shows a significant correlation to time of day. In particular, sessions started in the early morning are notably longer than sessions started in the afternoon or evening. For example, the fraction of sessions with duration below 90 minutes starting between 03:00 and 04:00 is 85%. However, this fraction is 93% for sessions starting between 13:00 and 14:00. A similar trend is observed for sessions started in the evening hours in North America, i.e. 11:00 to 12:00 at the measurement peer. For Asian peers the correlation between connected session duration and time of day is less obvious due to the relatively small sample size. Nevertheless, Figure 3.6 (c) indicates shorter average session durations in the peak period of Asian peers, i.e. between 13:00 and 14:00. We conclude that correlations between session duration and time of day are significant for generating synthetic workload.

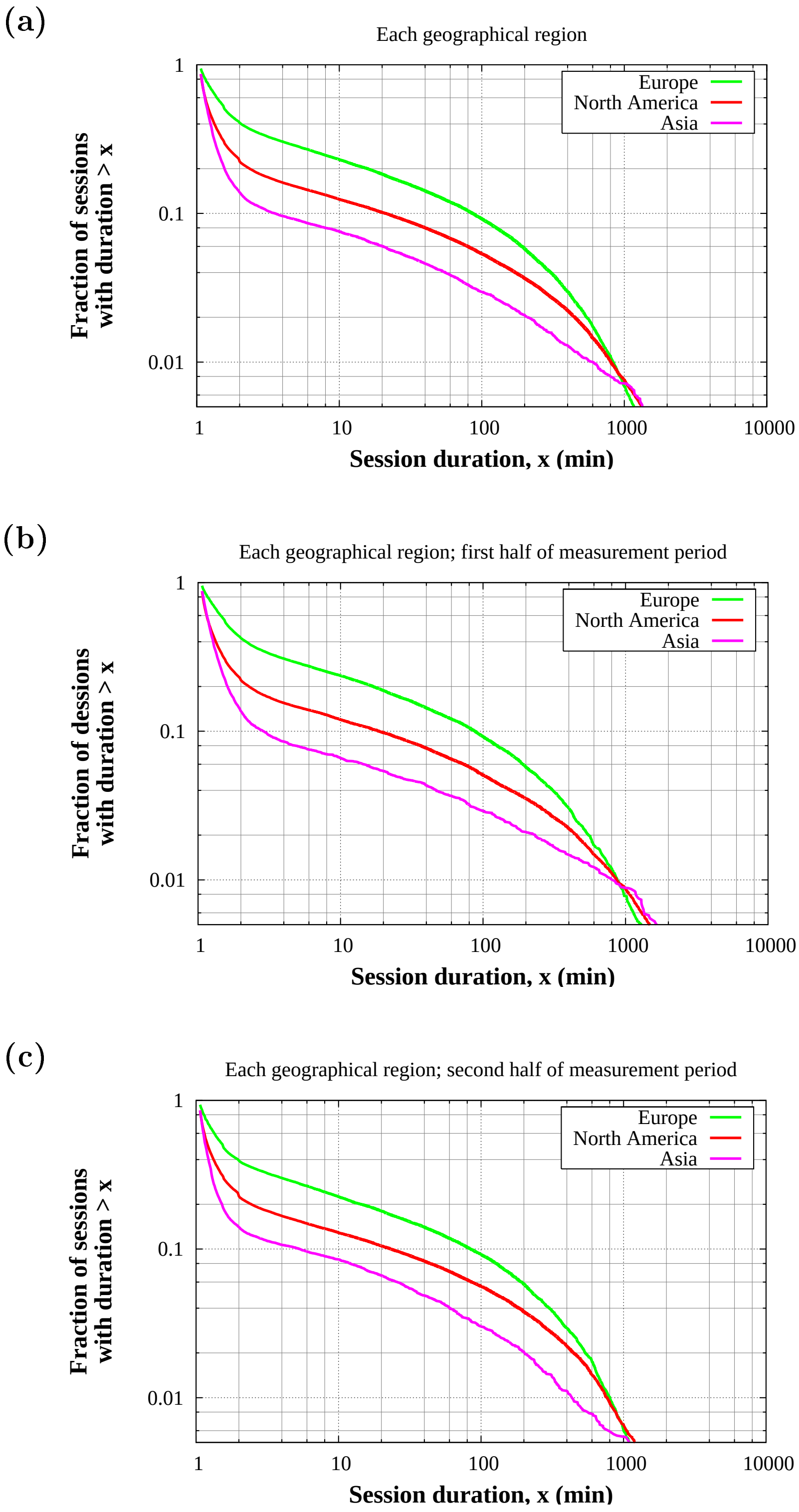


Figure 3.5: Distribution of connected session duration for passive peers broken down by geographical region

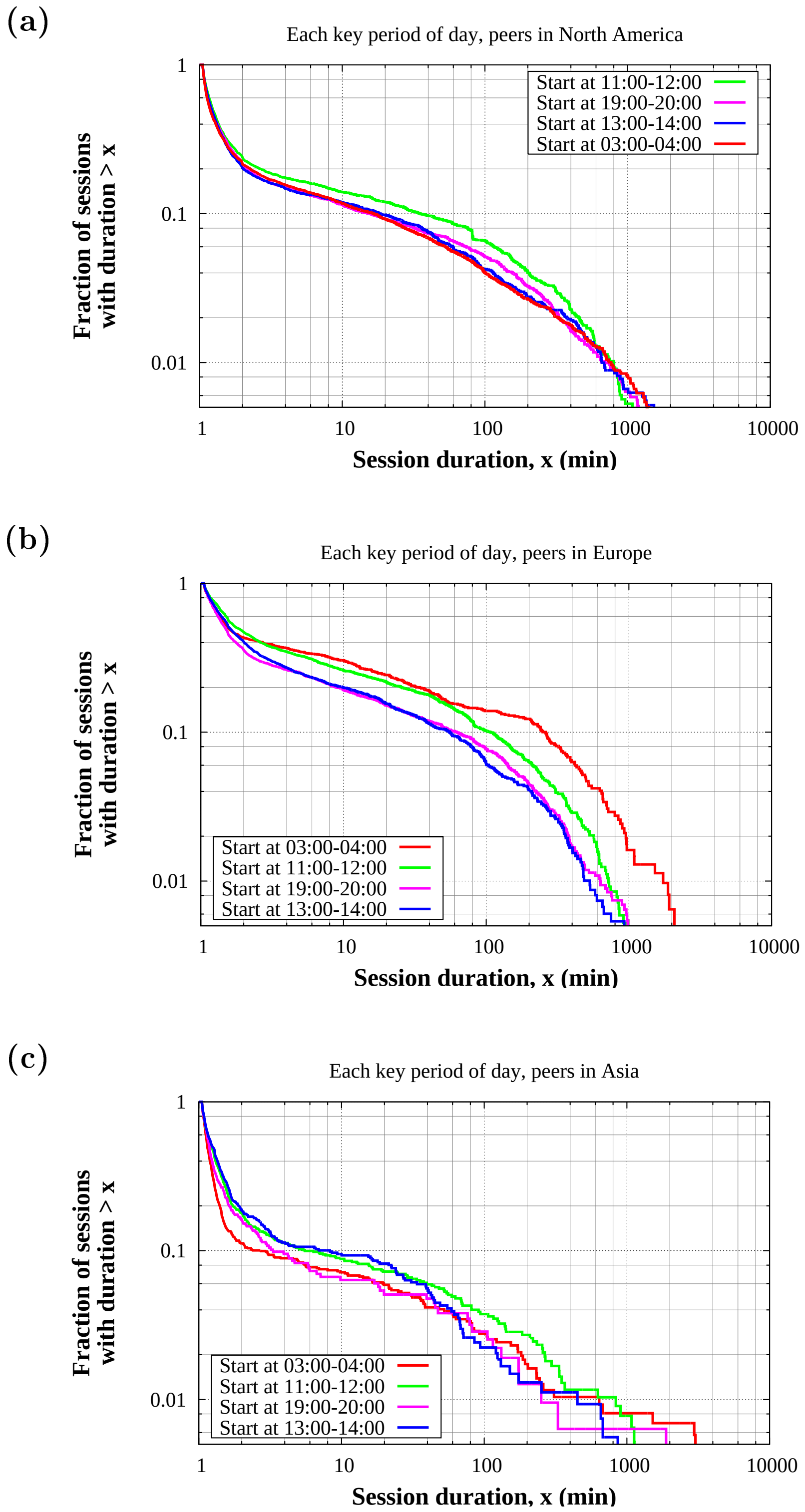


Figure 3.6: Distribution of connected session duration for passive peers broken down by time of day

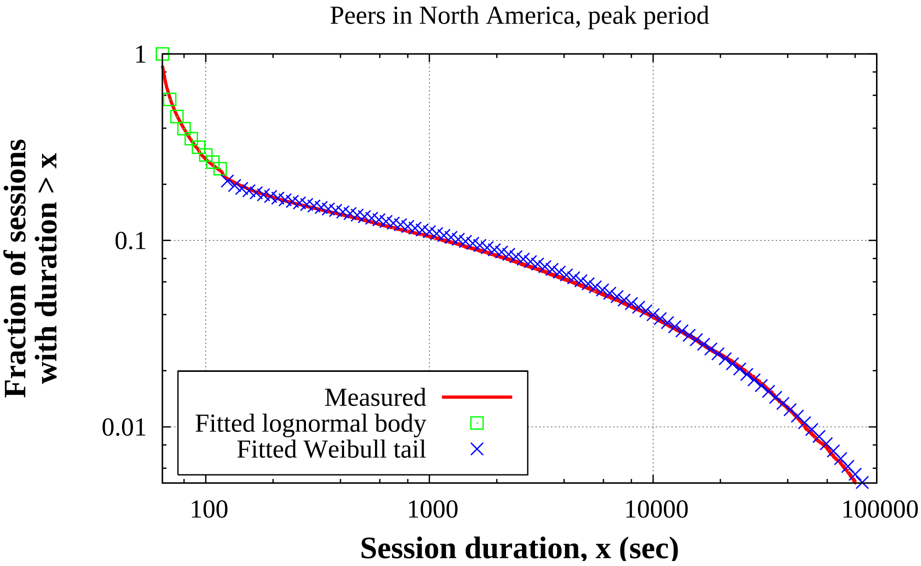


Figure 3.7: Fitting quality of connected session duration for passive peers

Period of the day and peer location		**Fitted distribution**	**Matched parameters**
Peak for North American peers	Body: 64-120 seconds	Lognormal	$\sigma = 2.690$, $\mu = 2.071$
	Tail: $>$ 120 seconds	Weibull	$\alpha = 0.3584$, $\lambda = 0.06529$
Non-peak for North American peers	Body: 64-120 seconds	Lognormal	$\sigma = 2.383$, $\mu = 2.201$
	Tail: $>$ 120 seconds	Weibull	$\alpha = 0.3662$, $\lambda = 0.05541$
Peak for European peers	Body: 64-150 seconds	Lognormal	$\sigma = 2.251$, $\mu = 3.374$
	Tail: $>$ 150 seconds	Weibull	$\alpha = 0.4406$, $\lambda = 0.03290$
Non-peak for European peers	Body: 64-150 seconds	Lognormal	$\sigma = 2.541$, $\mu = 3.895$
	Tail: $>$ 150 seconds	Weibull	$\alpha = 0.4457$, $\lambda = 0.02468$
Peak for Asian peers	Body: 64-120 seconds	Lognormal	$\sigma = 2.058$, $\mu = 2.217$
	Tail: $>$ 120 seconds	Lognormal	$\sigma = 2.579$, $\mu = 6.082$
Non-peak for Asian peers	Body: 64-104 seconds	Lognormal	$\sigma = 1.816$, $\mu = 1.852$
	Tail: $>$ 104 seconds	Weibull	$\alpha = 0.3093$, $\lambda = 0.1139$

Table 3.3: Model distributions and parameters for connected session duration of passive peers

Following the fitting process described in Section 2.3, the session duration for passive peers conditioned on geographical region and time of day can be well modeled by a hybrid distribution composed of a lognormal distribution for the body and a Weibull distribution for the tail. Figure 3.7 shows that such a hybrid distribution reasonably models the measured session durations. Note, that session durations of less than 64 seconds are not considered due to the filtering mechanism described in Section 3.3.3. That is, for fitting the model distributions to the measured data 64 seconds are subtracted from each measured session duration. Thus, a workload generator choosing the session duration according to the fitted distributions stated in Table 3.3 has to increase the results by 64 seconds.

3.4.5 Active Peer Session Characteristics

Connected session duration for active peers is a measure composed of the number of queries issued in a session, the time between the establishment of the connection and the sending of the first query, the time between sending two successive queries and the time after sending the last query until termination of the connection. Thus, in the following we characterize each of these measures separately.

Number of Queries per Session

Figure 3.8 (a) plots the CCDF of the number of queries per connected session broken down by geographical region. Recall that sessions with zero queries are passive and already considered in Sections 3.4.3 and 3.4.4. Thus, these sessions are neglected in the analysis of the number of queries per session. We observe a significant correlation between number of queries and geographical region. For instance, the fraction of peers that issue less than 5 queries is 93% for Asia, 86% for North America, and only 74% for Europe. 5% of the Asian peers issue 5 to 10 queries. In North America and in Europe, these fractions constitute 10% and 18%, respectively. Moreover, we find that sessions with more than 10 queries comprise 2% of the session in Asia, 4% in North America and 8% in Europe. As a consequence, we conclude that European peers issue significantly more queries in a session than peers from the other geographical regions. Again, considering multiple-day time periods by separating the first and the second half of the measurement period yields no significant difference in the corresponding curves.

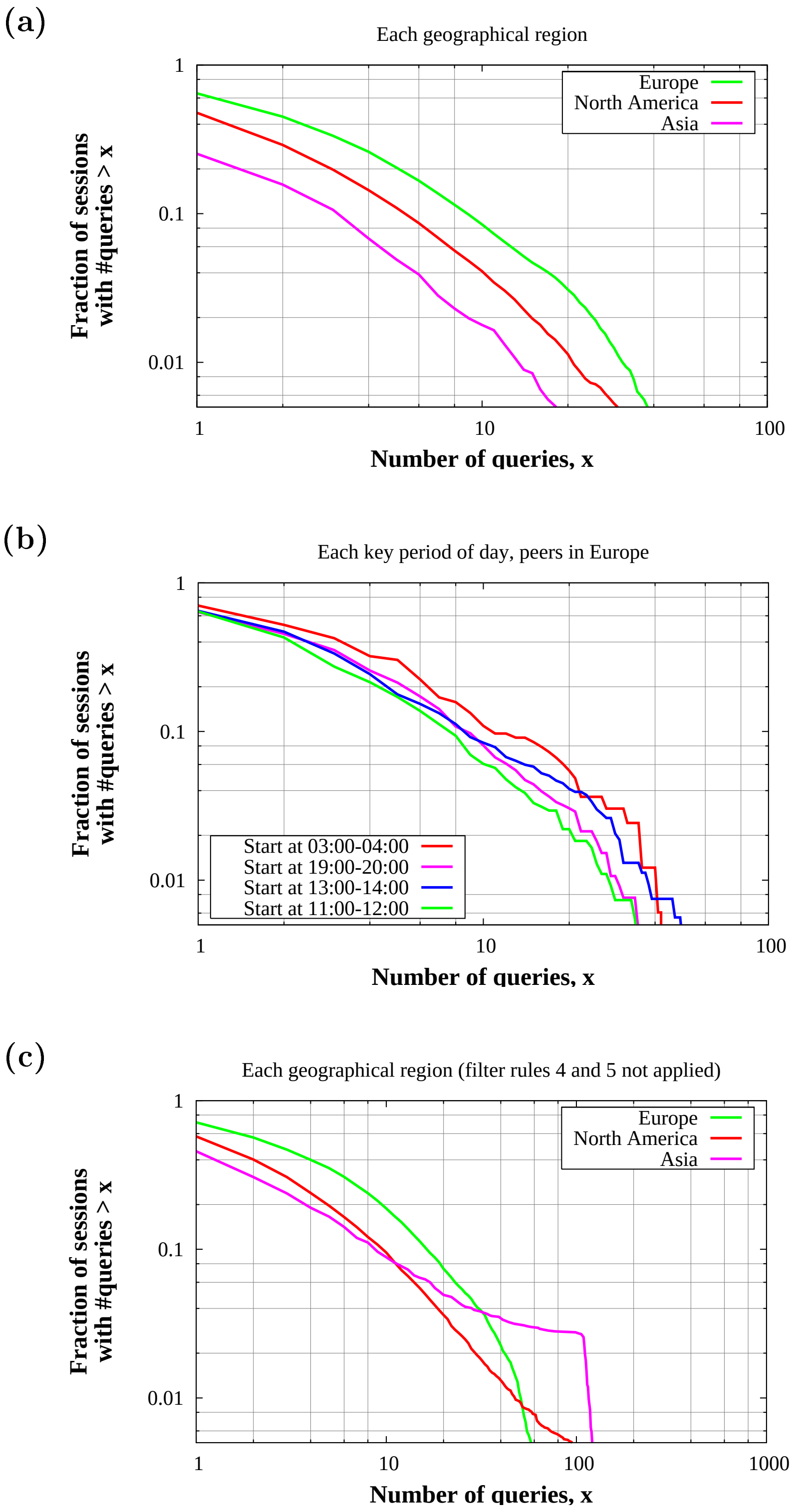


Figure 3.8: Distribution of number of queries per session

Geographical region	Fitted distribution	Matched parameters
North America	Lognormal	$\sigma = 1.306$, $\mu = 0.001$
Europe	Lognormal	$\sigma = 1.306$, $\mu = 0.520$
Asia	Weibull	$\alpha = 0.4772$, $\lambda = 1.360$

Table 3.4: Model distributions and parameters for number of queries per session

In a further experiment, we analyze the correlation between number of queries per session and time of day. Figure 3.8 (b) plots the CCDF of number of queries per session for sessions of European peers starting in each of the important periods identified in Section 3.4.2. We find that number of queries per session is roughly insensitive to session start time for 90% of the sessions from Europe. The same holds for sessions of peers from the other geographical regions.

In our measurement setup, we use the filter rules presented in Section 3.3.3 to discard all queries, which are automatically generated by Gnutella clients. The matching of rules 4 and 5 indicates that there are user-generated queries, which were issued before the peer connects to the measurement node. Thus, these queries contribute to the overall number of queries a user issues in his session, although we cannot determine the sending time due to the measurement setup. For completeness, we provide the distribution of the number of queries per session when filter rules 4 and 5 are not applied in Figure 3.8 (c). We observe that the number of queries without applying filter rules 4 and 5 most significantly changes for Asian peers. For Asian peers there is a large fraction of about 4% of sessions with more than 100 queries. Without analyzing this behavior in detail, we conjecture that the relatively large fraction of sessions with many automated but distinct queries are caused by a special feature of Asian Gnutella clients. The analysis in the remainder of this section is based on the number of queries with filter rules 4 and 5 applied.

Figure 3.9 graphically shows that the lognormal distribution is a suitable model for the number of queries per session for North American and European peers. The number of queries per session issued by Asian peers is better modeled by a Weibull distribution. The matched parameters of the model for the three most important geographical regions are stated in Table 3.4.

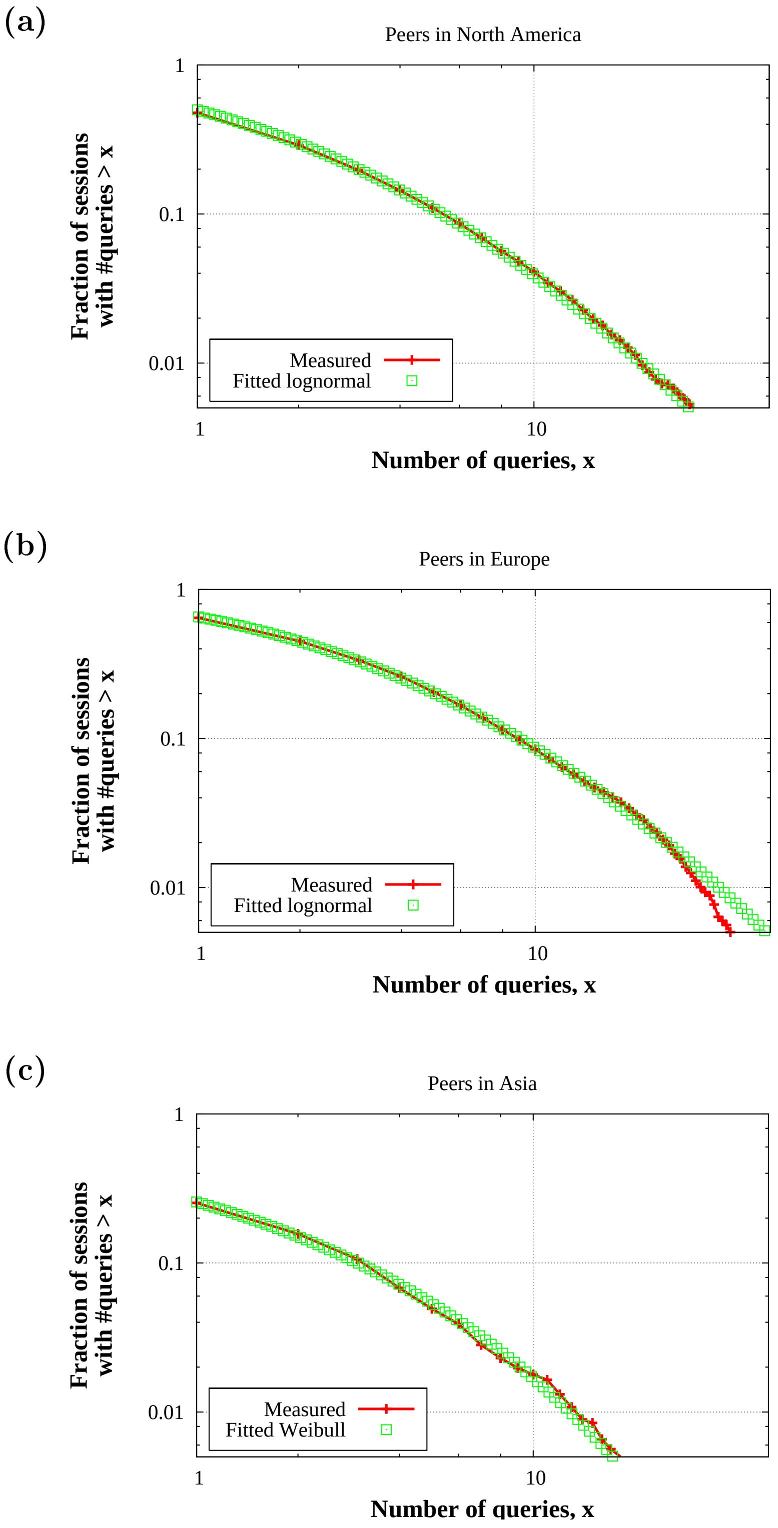


Figure 3.9: Fitting quality of number of queries per session

Time until First Query

We plot in Figure 3.10 (a) the CCDF of the time between connection establishment and sending of the first query broken down by geographical region. While the curves look very similar for North American and European peers, there is a significant difference for Asian peers. The first query within a session from Asian peers is issued within 10 seconds for 7% of the peers, whereas this fraction constitutes 20% for North America and Europe. The fraction of peers that issue a query within 30 seconds stays with 50% almost equal for all regions. Another 40% of the Asian peers issue the first query within 30 and 90 seconds. The same fraction of peers issues the first query within 30 and 1,000 seconds for Europe, indicating a significant correlation to geographical region. Furthermore, a fraction of 1% of the sessions started in North America and Europe issue the first query after 33,000 seconds, i.e. after more than 9 hours.

To analyze correlations between time until first query and number of queries issued in a session, Figure 3.11 plots the percentile curves for time until first query conditioned on the number of queries issued in the session. We observe for each of the important geographical regions that the 20th and 50th percentiles are almost equal for all number of queries. That is, at least for the half of the sessions with small time until first query, there is no correlation to the number of queries. In contrast, for the sessions with longer time until first query there is correlation, which is very different for the three geographical regions. For North American peers time until first query shows an increasing trend with increasing number of queries. We can identify three noticeable regions of the number of queries: less than 3 queries, exactly 3 queries and more than 3 queries. European peers behave very different. Here, significant differences in the 80^{th} percentile can be identified for sessions with less or exactly 8 queries and for sessions with more than 8 queries. For Asian peers there is no clear trend, which we attribute to the small sample set in the measurements. Thus, for Asian peers we neglect potential correlations between time until first query and number of queries per session.

Figure 3.10 (b) plots the CCDF of time until first query for North American peers broken down by the three regions of the number of queries identified above. This figure underlines that the conditional distributions are equal for 50% of the sessions with an early first query, while there is a significant difference for the 50% of the session with late first query. In particular, in 90% of the sessions with less than 3 queries the first query is issued before 200 seconds, in the sessions

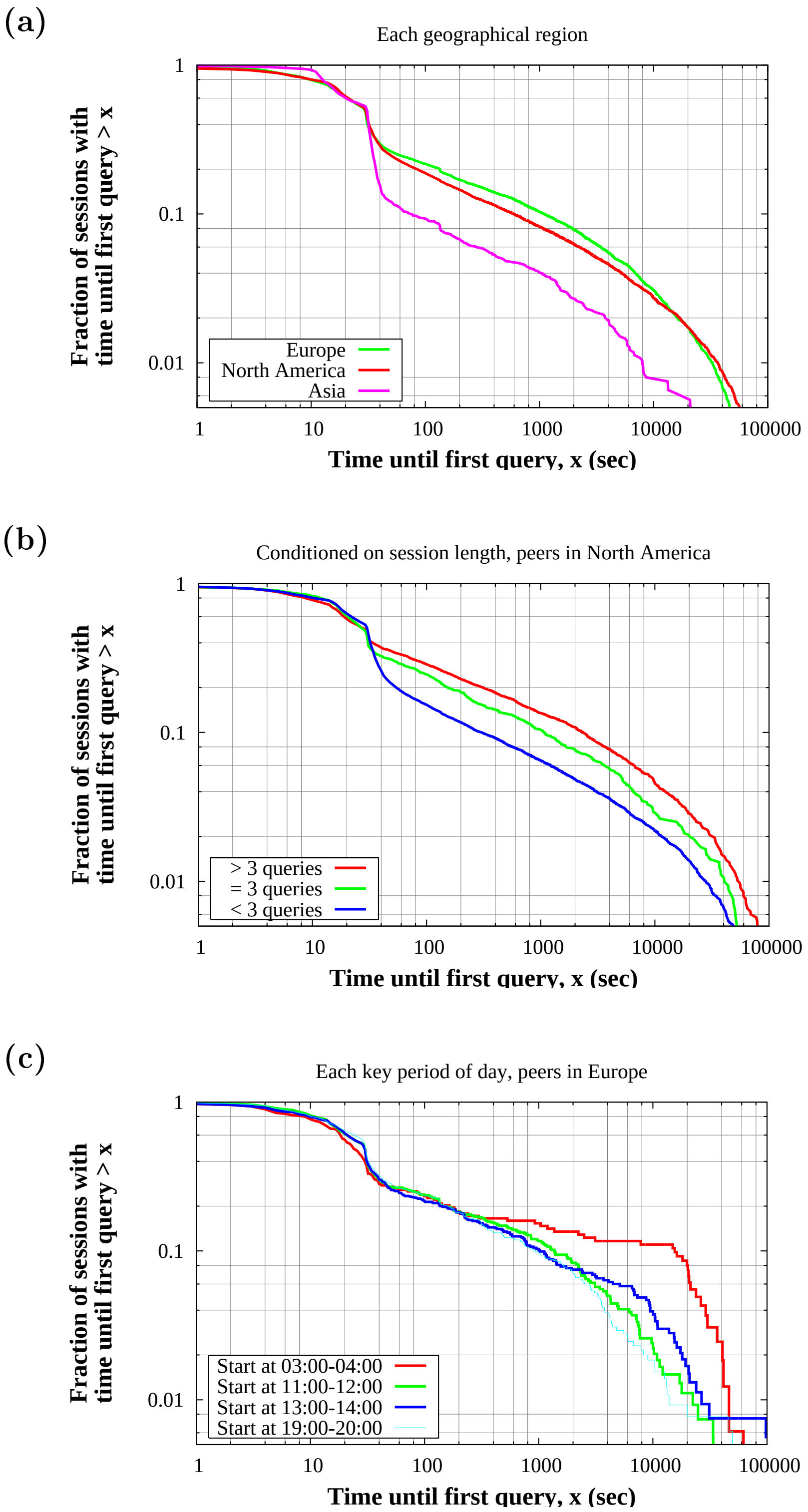


Figure 3.10: Distribution of time until first query

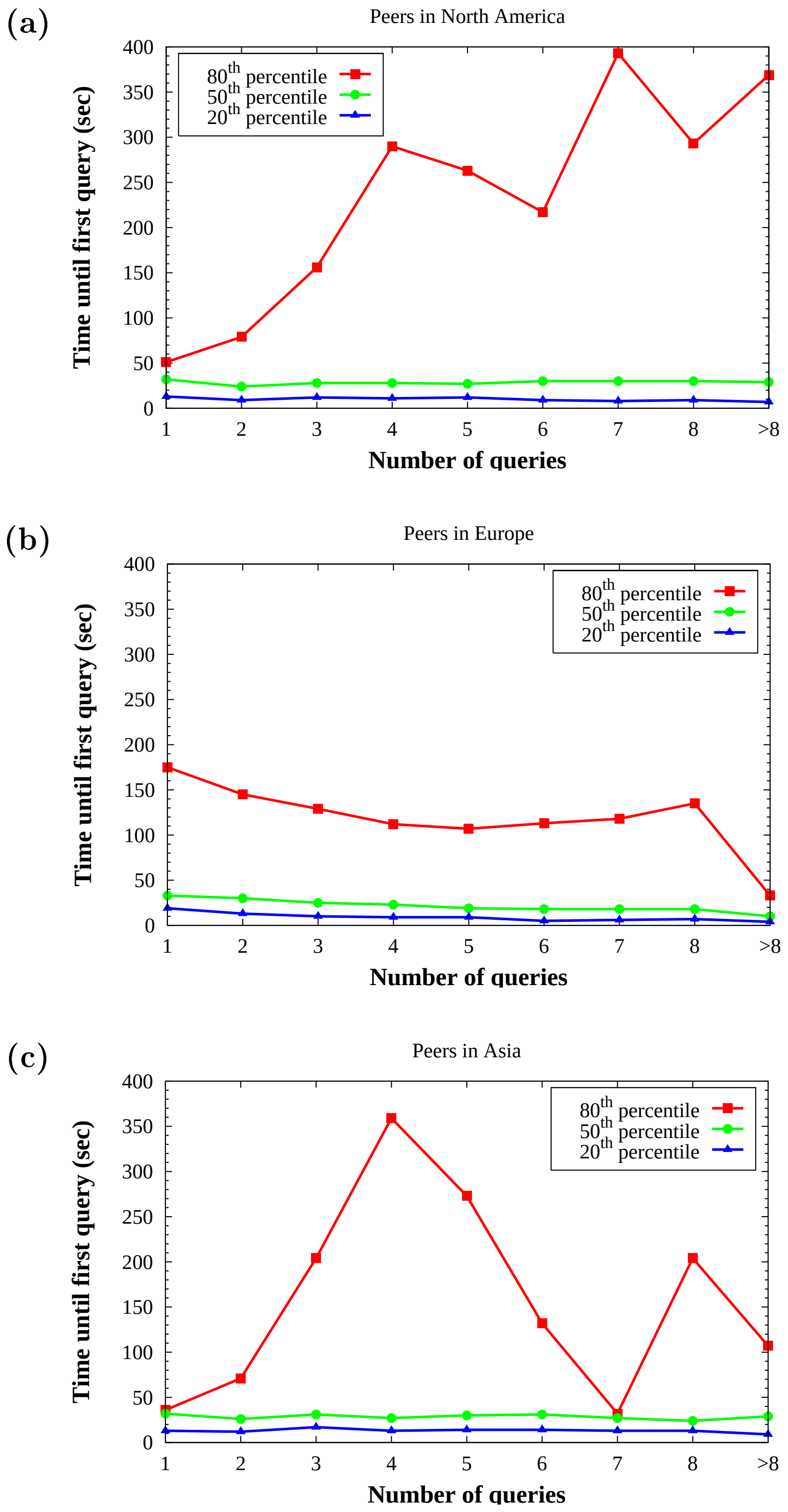


Figure 3.11: Correlation between time until first query and number of queries

with exactly 3 queries before 1,000 seconds and in sessions with more than 3 queries before 2,000 seconds. We conclude from Figure 3.11 and Figure 3.10 (b) that time until first query is correlated with number of queries per session for North American and European peers.

Figure 3.10 (c) plots the CCDF of time until first query for the important daily time periods of European peers identified above. We find that in a significant fraction of sessions started in the non-peak hours the first query is sent 10,000 seconds and more after session start. These fractions constitute 10% for Europe. The same trend can be observed for the other geographical regions. We conclude from Figure 3.10 (c) that there is a significant correlation between time of day and time until first query. To capture this correlation, the workload model distinguishes between sessions starting in peak and in non-peak-hours.

The time until the first query conditioned on geographical region, time of day and number of queries per session can be modeled by a hybrid distribution composed of a Weibull distribution and a lognormal distribution as shown in Figure 3.12. Note that depending on the geographical region and the number of queries per session different combinations of Weibull and lognormal distributions fit best to the measured data. The parameters for the conditional distributions are presented in Table 3.5. Since the number of measured sessions from Asian peers is too small to identify significant correlation to number of queries per session Table 3.5 provides the fitted distributions for Asian peers without distinguishing different number of queries per session.

Query Interarrival Time

We denote the time between issuing two successive queries as query interarrival time. Figure 3.13 (a) plots the CCDF of query interarrival time broken down by geographical region. This figure shows that queries generated by European peers have shorter interarrival times than queries from the other two regions. For instance, the fraction of interarrival times below 100 seconds constitutes 90% for Europe, while it is 83% for Asia and 75% for North America. We conclude from Figure 3.13 (a) that query interarrival time shows a significant correlation to geographical region.

The analysis of the correlation between query interarrival time and number of queries per session reveals some interesting results. Percentile curves in Figure 3.14(a) show that for North American peers there seems to be a slight trend

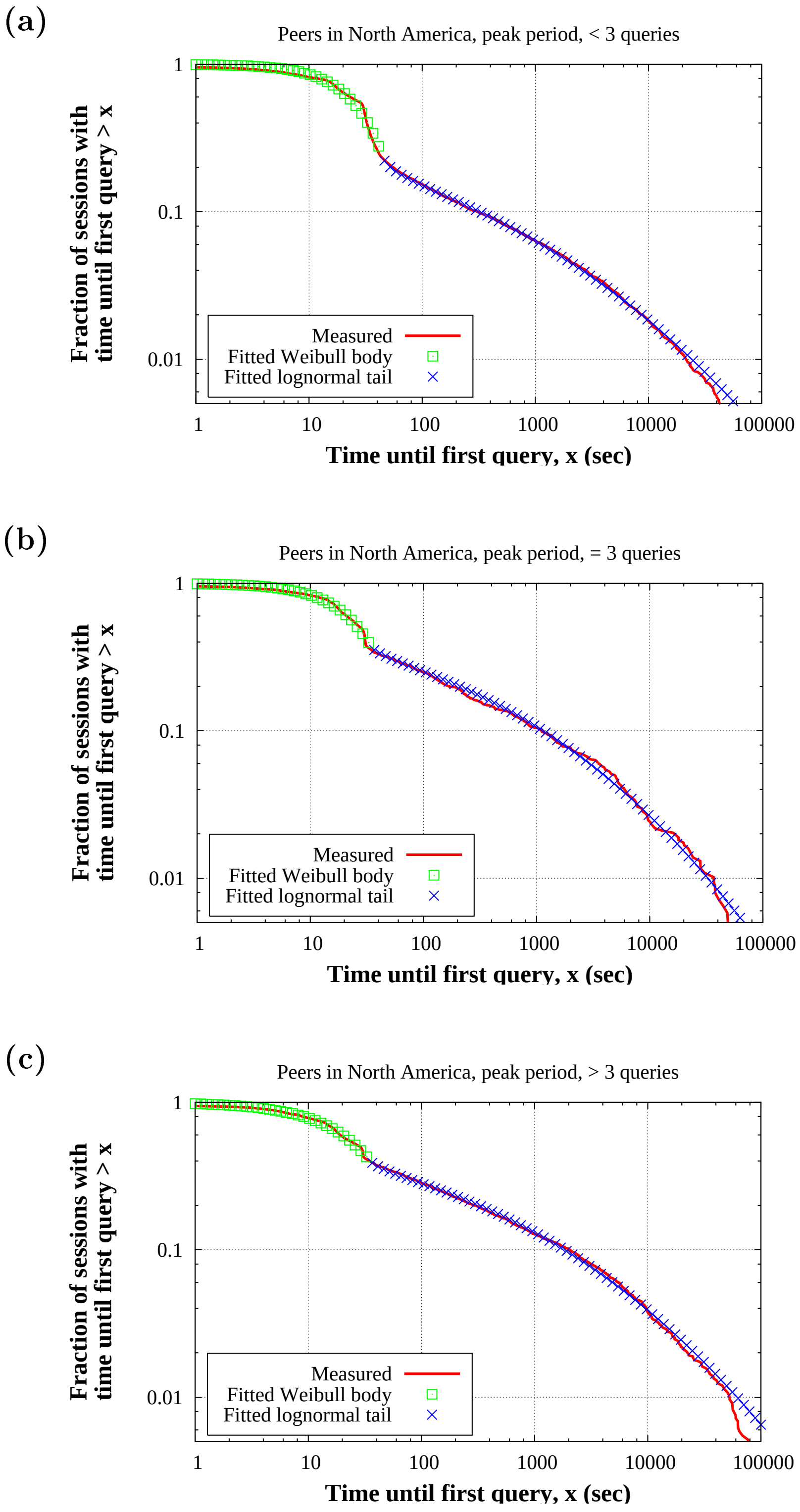


Figure 3.12: Fitting quality of time until first query

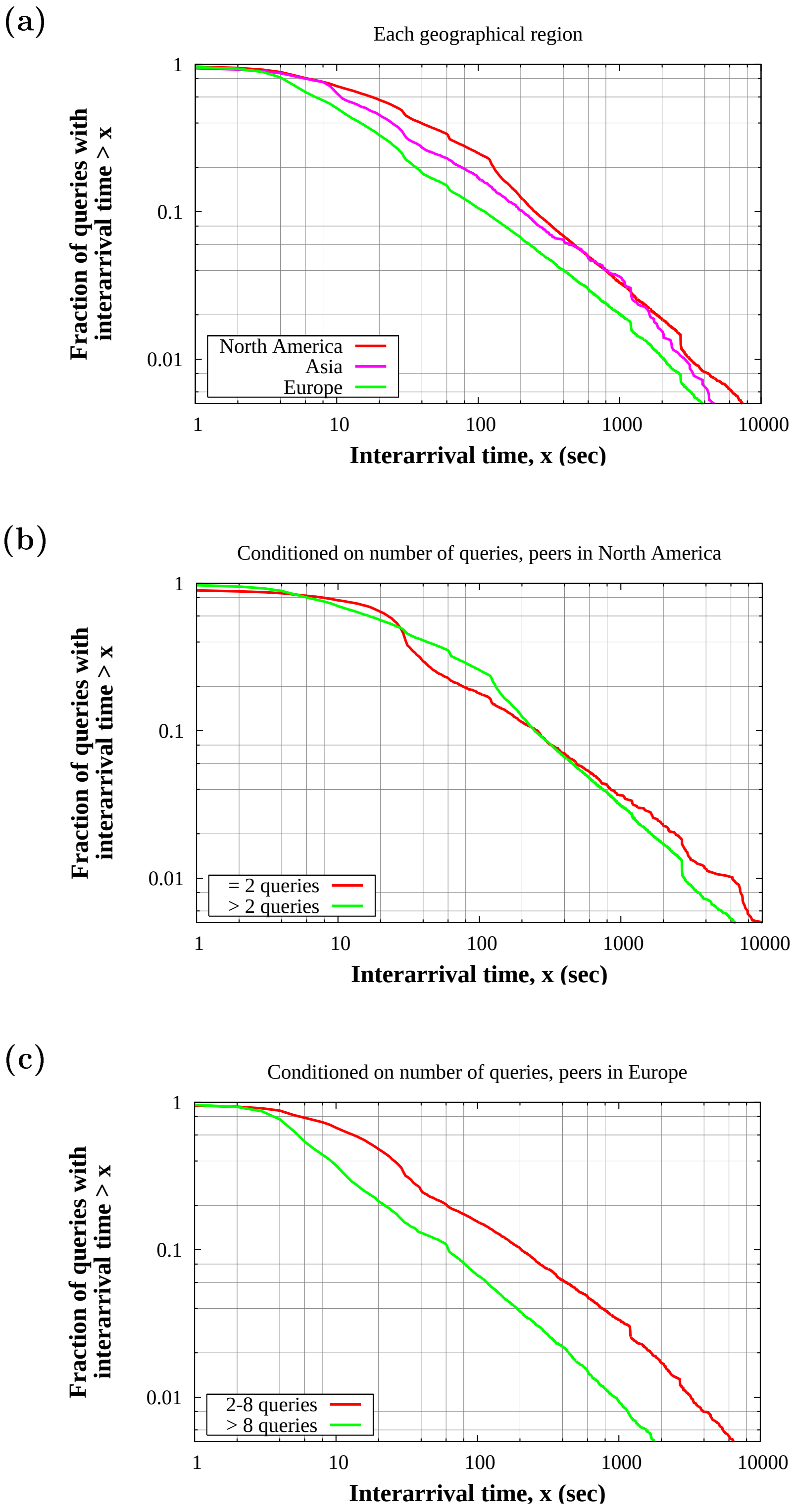


Figure 3.13: Distribution of query interarrival time

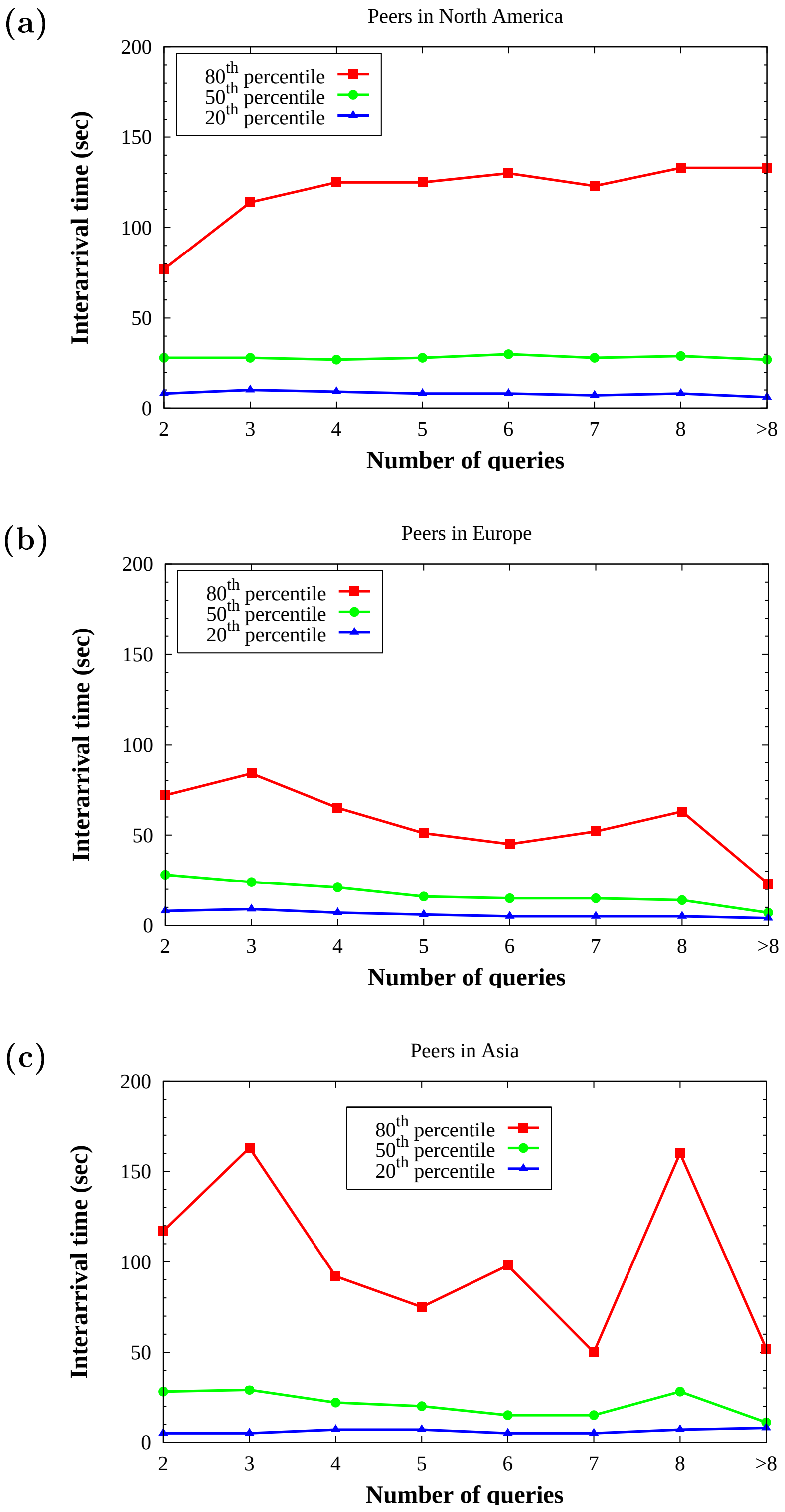


Figure 3.14: Correlation between query interarrival time and number of queries

Period of the day and peer location	**Number of queries per session**		**Fitted distribution**	**Matched parameters**
Peak for North American peers	$<$ 3 queries	Body: 0-45 seconds	Weibull	$\alpha = 1.477$, $\lambda = 0.005247$
		Tail: $>$ 45 seconds	Lognormal	$\sigma = 2.858$, $\mu = 5.171$
	$=$ 3 queries	Body: 0-35 seconds	Weibull	$\alpha = 1.364$, $\lambda = 0.007930$
		Tail: $>$ 35 seconds	Lognormal	$\sigma = 2.562$, $\mu = 5.476$
	$>$ 3 queries	Body: 0-35 seconds	Weibull	$\alpha = 1.041$, $\lambda = 0.02252$
		Tail: $>$ 35 seconds	Lognormal	$\sigma = 2.748$, $\mu = 5.634$
Non-peak for North American peers	$<$ 3 queries	Body: 0 - 50 seconds	Weibull	$\alpha = 1.156$, $\lambda = 0.01796$
		Tail: $>$ 50 seconds	Lognormal	$\sigma = 3.150$, $\mu = 5.647$
	$=$ 3 queries	Body: 0 - 35 seconds	Weibull	$\alpha = 1.240$, $\lambda = 0.01308$
		Tail: $>$ 35 seconds	Lognormal	$\sigma = 2.849$, $\mu = 5.394$
	$>$ 3 queries	Body: 0 - 35 seconds	Weibull	$\alpha = 0.9892$, $\lambda = 0.02906$
		Tail: $>$ 35 seconds	Lognormal	$\sigma = 2.996$, $\mu = 6.404$
Peak for European peers	$<=$ 8 queries	Body: 0-40 seconds	Weibull	$\alpha = 1.428$, $\lambda = 0.005849$
		Tail: $>$ 40 seconds	Weibull	$\alpha = 0.3788$, $\lambda = 0.08184$
	$>$ 8 queries	Body: 0-40 seconds	Lognormal	$\sigma = 1.393$, $\mu = 2.330$
		Tail: $>$ 40 seconds	Lognormal	$\sigma = 2.418$, $\mu = 6.069$
Non-peak for European peers	$<=$ 8 queries	Body: 0 - 35 seconds	Weibull	$\alpha = 1.355$, $\lambda = 0.007818$
		Tail: $>$ 35 seconds	Weibull	$\alpha = 0.3032$, $\lambda = 0.1240$
	$>$ 8 queries	Body: 0 - 35 seconds	Lognormal	$\sigma = 1.320$, $\mu = 2.214$
		Tail: $>$ 35 seconds	Weibull	$\alpha = 0.3938$, $\lambda = 0.04900$
Asian peers		Body: 0-35 seconds	Weibull	$\alpha = 1.872$, $\lambda = 0.001604$
		Tail: $>$ 35 seconds	Lognormal	$\sigma = 2.666$, $\mu = 5.416$

Table 3.5: Model distributions and parameters for time until first query

towards longer interarrival times in sessions with more queries, indicated by the increasing 80th percentile curve. This trend is most significant for number of queries of exactly 2 compared with number of queries of more than 2. In contrast, for European peers the query interarrival times tend to decrease with increasing number of queries per session. The most obvious difference in interarrival times is between sessions with less or equal 8 queries and sessions with more than 8 queries. As in the previous discussion of the time until first query the percentile plots for Asian peers show no clear correlation. Thus, we again neglect a potential correlation.

For a more detailed analysis Figures 3.13 (b) and 3.13 (c) plot the CCDF of query interarrival time conditioned on the number of queries for North American and European peers, respectively. Whereas there is no significant correlation between these two measures for North American peers, sessions of European peers with many queries have smaller interarrival times than sessions with few queries. This indicates that there is a difference in the environment of North American and European peers like e.g. the prizing model of the Internet service providers. Due to this difference, sessions of North American peers with many queries tend to connect for a longer period compared to similar sessions of European peers. We conclude that query interarrival time has to be conditioned on number of queries per session for European peers but not for North American peers.

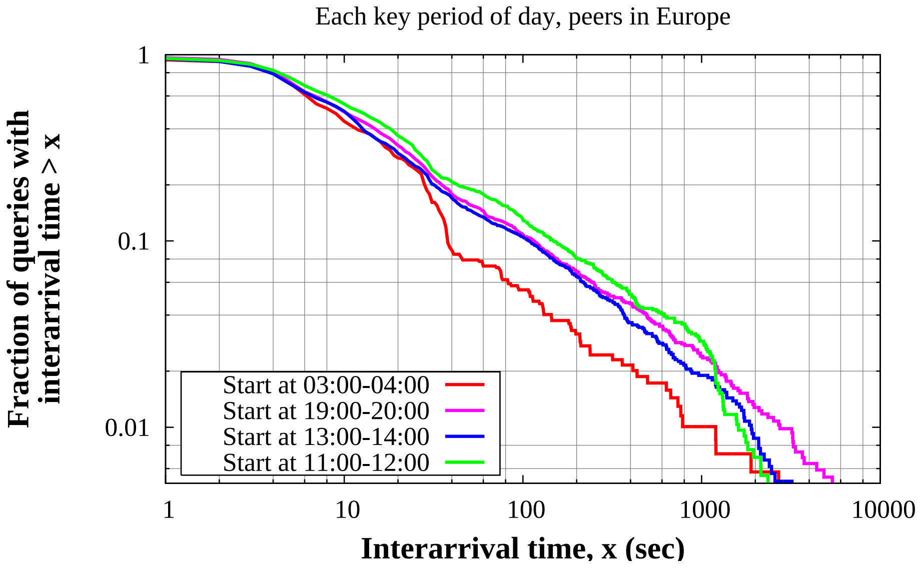


Figure 3.15: Query interarrival time conditioned on time of day

To analyze the correlation to time of day, Figure 3.15 plots the CCDF of query interarrival time broken down to the important daily time periods for European peers. It shows that queries issued in peak hours (all daily periods except 03:00-04:00) have longer interarrival times than queries issued in non-peak hours. For example, 94% of the queries issued in Europe between 3:00 and 4:00 have an interarrival time below 100 seconds, while this fraction is only 85% for sessions starting between 11:00 and 12:00. Results for the other geographical regions are identical. We conclude that query interarrival time shows a significant correlation to time of day.

As before, we provide a ready-to-use model for the conditional distributions of query interarrival time. Figure 3.16 shows that a hybrid distribution composed of a lognormal body and a Pareto tail matches well to the measured query interarrival times. We conclude from this figure that the distribution of query interarrival times is heavy-tailed. The parameters for the conditional distributions are summarized in Table 3.6.

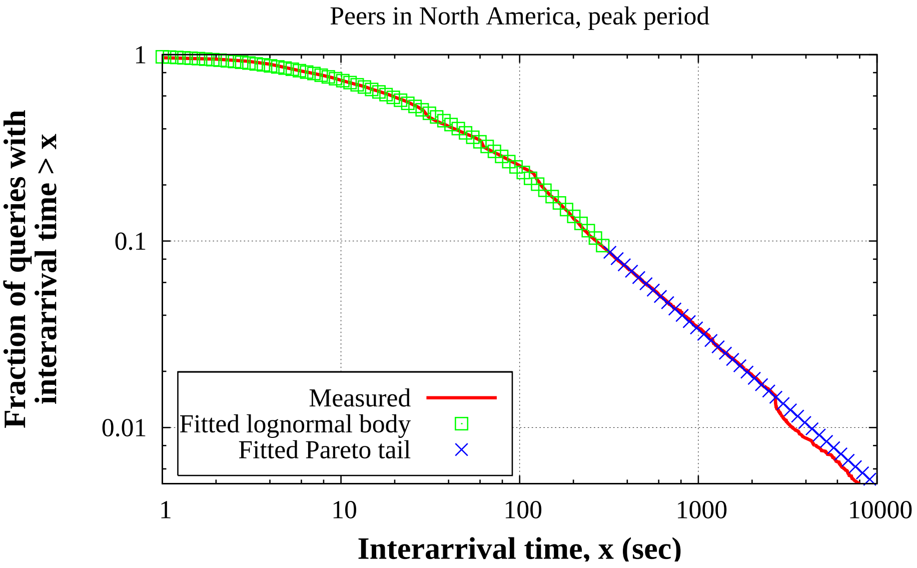


Figure 3.16: Fitting quality of query interarrival time

Period of the day and peer location	Number of queries per session		Fitted distribution	Matched parameters
Peak for North American peers		Body: $\leq$ 300 seconds	Lognormal	$\sigma = 1.752$, $\mu = 3.375$
		Tail: $>$ 300 seconds	Pareto	$\alpha = 0.8368$, $k = 300$
Non-peak for North American peers		Body: $\leq$ 300 seconds	Lognormal	$\sigma = 1.830$, $\mu = 3.053$
		Tail: $>$ 300 seconds	Pareto	$\alpha = 0.8117$, $k = 300$
Peak for European peers	2–8 queries	Body: $\leq$ 30 seconds	Lognormal	$\sigma = 1.403$, $\mu = 2.953$
		Tail: $>$ 30 seconds	Pareto	$\alpha = 0.6451$, $k = 30$
	$>$ 8 queries	Body: $\leq$ 30 seconds	Lognormal	$\sigma = 1.194$, $\mu = 2.051$
		Tail: $>$ 30 seconds	Pareto	$\alpha = 0.7128$, $k = 30$
Non-peak for European peers	2–8 queries	Body: $\leq$ 30 seconds	Lognormal	$\sigma = 1.2715$, $\mu = 2.827$
		Tail: $>$ 30 seconds	Pareto	$\alpha = 0.6890$, $k = 30$
	$>$ 8 queries	Body: $\leq$ 30 seconds	Lognormal	$\sigma = 1.121$, $\mu = 1.962$
		Tail: $>$ 30 seconds	Pareto	$\alpha = 0.7962$, $k = 30$
Peak for Asian peers		Body: $\leq$ 30 seconds	Lognormal	$\sigma = 1.406$, $\mu = 2.594$
		Tail: $>$ 30 seconds	Pareto	$\alpha = 0.4564$, $k = 30$
Non-peak for Asian peers		Body: $\leq$ 30 seconds	Lognormal	$\sigma = 1.337$, $\mu = 2.859$
		Tail: $>$ 30 seconds	Pareto	$\alpha = 0.6398$, $k = 30$

Table 3.6: Model distributions and parameters for query interarrival time

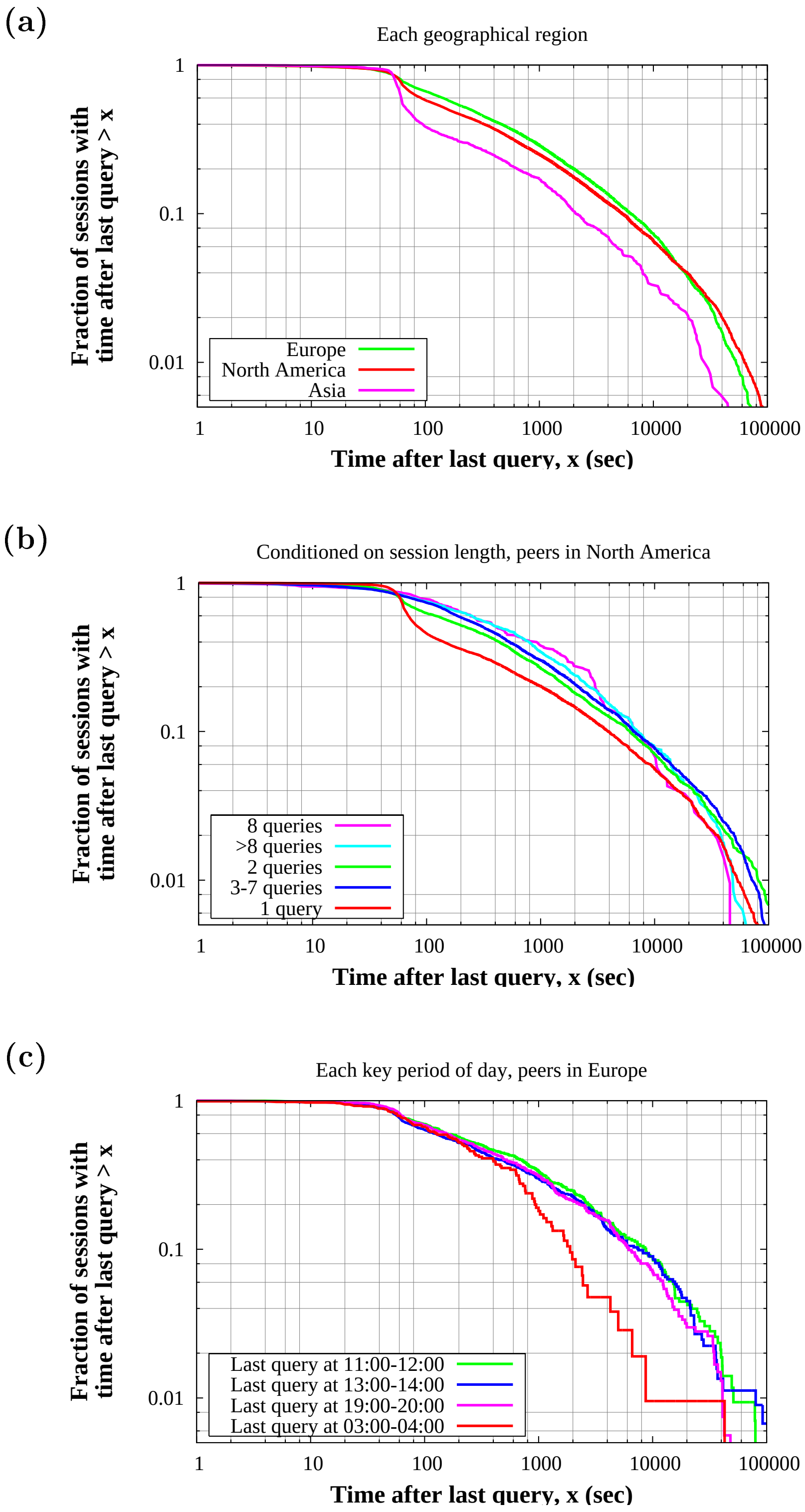


Figure 3.17: Distribution of time after last query

Time after Last Query

Figure 3.17 (a) plots the CCDF of the time between sending the last query and termination of the connection broken down by geographical region. The figure shows that only a very small fraction of peers close the connection in less than 12 seconds after the last query. Furthermore, the distributions are very similar for North American and European peers, while Asian peers tend to close sessions much faster. For instance, the fraction of sessions with a time after last query of more than 1000 seconds is 30% and 25% for Europe and North America, while it is only 18% for Asia. Recall that the time after last query may be overestimated by approximately 30 seconds due to the measurement setup. We conclude that there is a significant correlation between time after last query and geographical region.

As first indication for the correlation between time after last query and number of queries per session Figure 3.18 shows the corresponding percentile plots for North American, European, and Asian peers. The plots show a slight positive correlation for North American and Asian peers, whereas for European peers there is no significant correlation. From the figure we identify the following distinguishing ranges for the number of queries: a single query, 2 queries, between 3 and 7 queries, 8 queries and more than 8 queries. The CCDF of time after last query conditioned on number of queries for North American peers is shown in Figure 3.17 (b). We observe the smallest and largest values for the time after the last query for sessions with a single query, and with 8 and more queries, respectively. Furthermore, the conditional distributions for 2 queries and 3 to 7 queries are identical for 99% of the sessions, similar to the curves for exactly 8 and more than 8 queries. Combining both distributions of these pairs, we observe a positive correlation between time after last query and number of queries per session for 90% of the sessions. We conclude from Figures 3.18 and 3.17 (b) that the distribution of time after last query must be conditioned on number of queries per session for North American and Asian peers.

Analyzing the correlation to time of day for European peers in Figure 3.17 (c), we find that sessions sending the last query in the non-peak hours have a shorter time after last query than sessions sending the last query in peak hours. This trend is most noticeable in Europe, where the time after the last query for sessions sending the last query between 03:00 and 04:00 is below 10,000 seconds for more than 99% of the sessions, while it is below 91% of the sessions sending

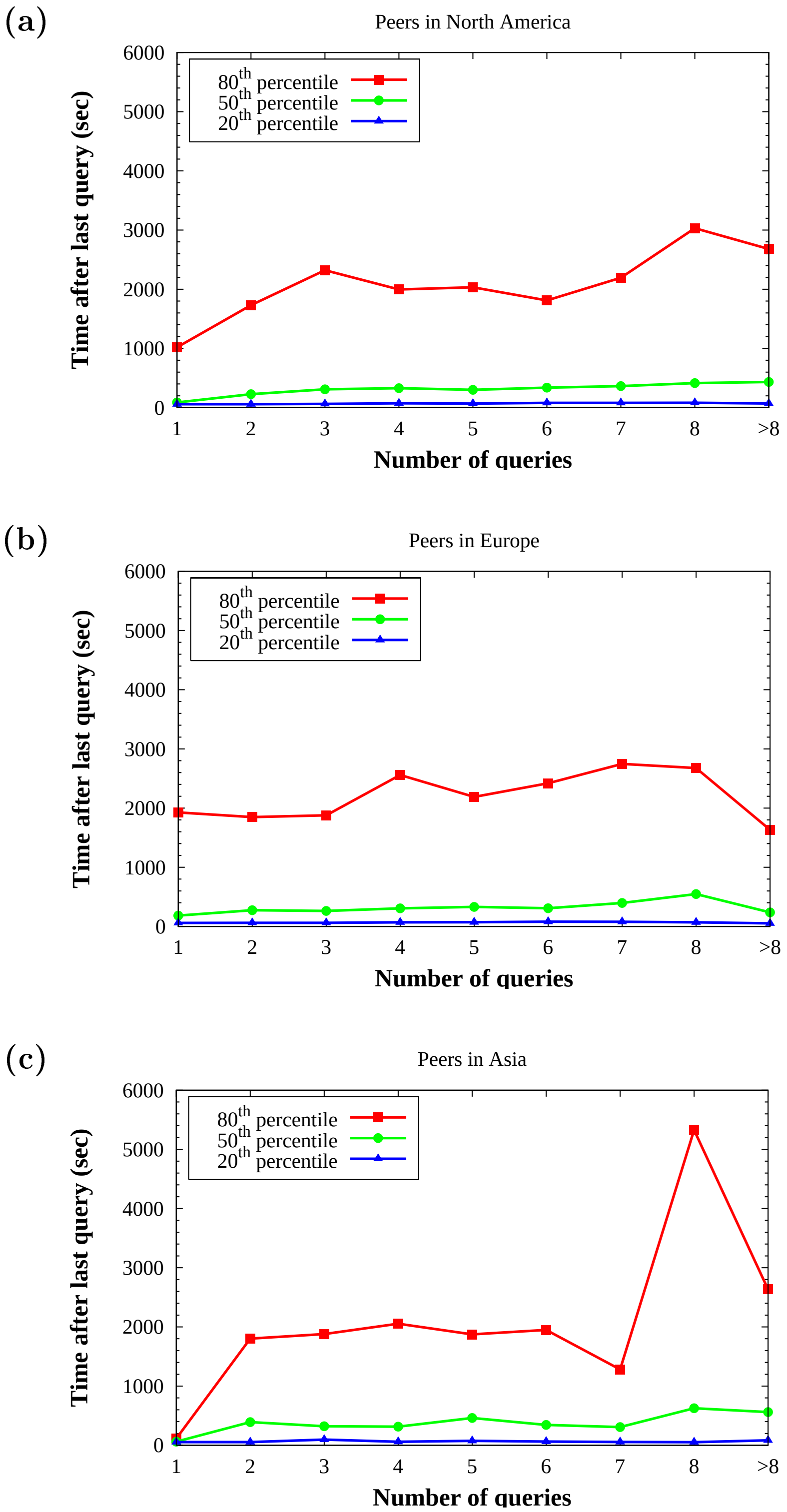


Figure 3.18: Correlation between time after last query and number of queries

Period of the day and peer location	**Number of queries per session**		**Fitted distribution**	**Matched parameters**
Peak for North American peers	1 query	Body: 0-70 seconds	Lognormal	$\sigma = 0.3395$, $\mu = 4.330$
		Tail: $>$ 70 seconds	Lognormal	$\sigma = 2.658$, $\mu = 5.668$
	2-7 queries		Lognormal	$\sigma = 2.259$, $\mu = 5.686$
	$>$ 7 queries		Lognormal	$\sigma = 2.145$, $\mu = 6.107$
Non-peak for North American peers	1 query	Body: 0-70 seconds	Lognormal	$\sigma = 0.2931$, $\mu = 4.277$
		Tail: $>$ 70 seconds	Weibull	$\alpha = 0.3299$, $\lambda = 0.1067$
	2-7 queries		Lognormal	$\sigma = 2.156$, $\mu = 5.672$
	$>$ 7 queries		Lognormal	$\sigma = 2.286$, $\mu = 6.036$
Peak for European peers			Lognormal	$\sigma = 2.270$, $\mu = 5.577$
Non-peak for European peers			Lognormal	$\sigma = 2.445$, $\mu = 5.756$
Peak for Asian peers	1 query	Body: 0-60 seconds	Lognormal	$\sigma = 0.1575$, $\mu = 4.126$
		Tail: $>$ 60 seconds	Lognormal	$\sigma = 2.730$, $\mu = 3.177$
	2-7 queries		Lognormal	$\sigma = 2.122$, $\mu = 5.948$
	$>$ 7 queries		Lognormal	$\sigma = 2.051$, $\mu = 5.785$
Non-peak for Asian peers	1 query	Body: 0-60 seconds	Lognormal	$\sigma = 0.1685$, $\mu = 4.118$
		Tail: $>$ 60 seconds	Lognormal	$\sigma = 3.367$, $\mu = 3.144$
	2-7 queries		Lognormal	$\sigma = 2.109$, $\mu = 5.850$
	$>$ 7 queries		Lognormal	$\sigma = 2.043$, $\mu = 6.301$

Table 3.7: Model distributions and parameters for time after last query

the last query at other times. For North American peers we observe the same trend. We conclude from Figure 3.17 (c) that time after last query is correlated to time of day.

The time after the last query conditioned on geographical region, time of day and number of queries per session can be well modeled by combinations of lognormal and Weibull distributions. As shown in Figure 3.19 (a) the distribution for sessions with a single query of North American peers at the peak time of day is best modeled by a combination of two lognormal distributions for the body and the tail, respectively. For sessions with more than a single query, the distribution of time after last query follow a lognormal distribution as shown in Figure 3.19 (b). As before all distributional models and the corresponding parameters are provided in Table 3.7.

3.4.6 Query Popularity Distribution

Since users interests will change over time, we expect that the set of queries that are popular will change within the measurement period. To confirm this assumption, we illustrate the hot set drift, i.e. the drift in the most popular queries [MEW00]. For illustration, we determine the number of the top ten queries on day n that are found among the top N queries on the subsequent day, for $N = 10$, 20 and 100. Furthermore, we perform the same experiment for the queries with

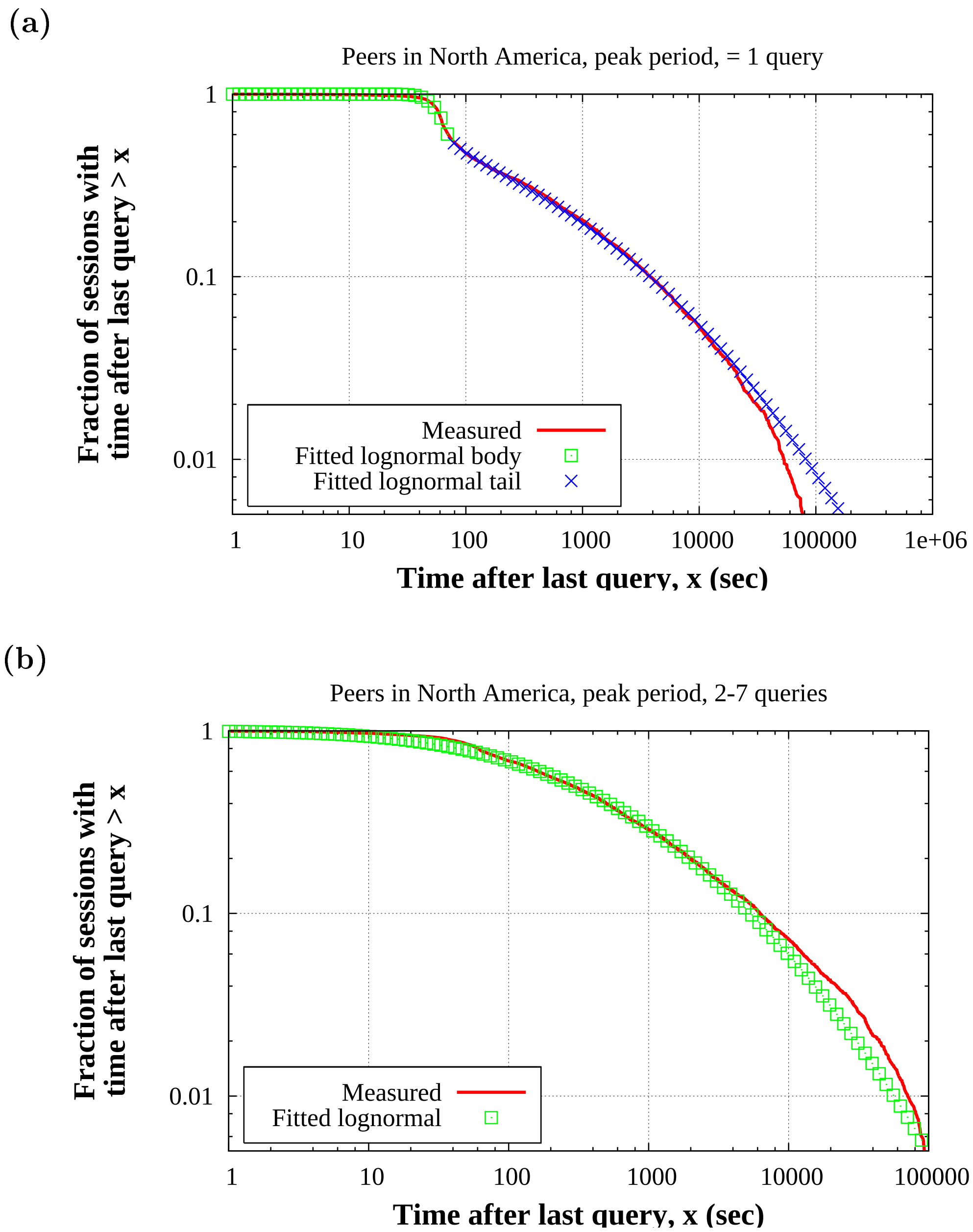


Figure 3.19: Fitting quality of time after last query

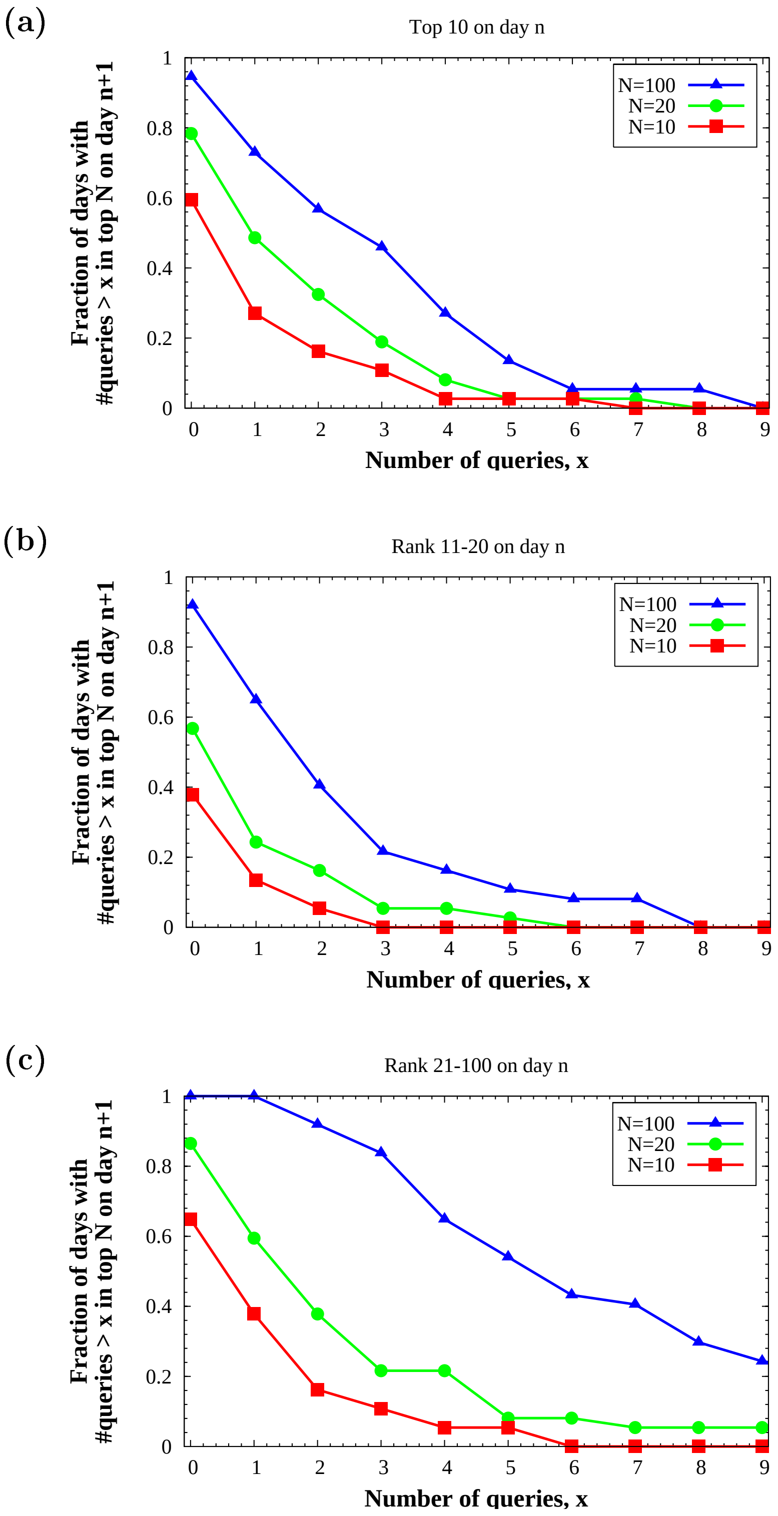


Figure 3.20: Drift in query popularity (peers in North America)

Measure	**4-day period**	**2-day period**	**1-day period**
Number of different queries from North American peers	6106	3588	1990
Number of different queries from European peers	5382	3729	1934
Number of different queries from Asian peers	776	299	153
Number of queries in intersection set between North American and European peers	323	114	56
Number of queries in intersection set between North American and Asian peers	41	15	5
Number of queries in intersection set between European and Asian peers	28	10	5
Number of queries in intersection set between North American, European and Asian peers	17	4	2

Table 3.8: Query class sizes

ranks 11-20 and 21-100 on day n, respectively. Figure 3.20 plots the CCDF of the observed distributions for North American peers. The figure shows that for about 70% of the days the number of top 10 queries that are found in the top 100 on the subsequent day is not larger than 4, indicating a significant hot set drift. Even the top 100 queries change significantly from day to day, as Figure 3.20 (c) illustrates. We conclude from Figure 3.20 that the query popularity distribution cannot be calculated over the entire trace, since the hot set drift must be considered.

In addition to temporal drift, we conjecture that the query popularity distribution depends on the geographical location of peers. To confirm this, we determine the set of distinct queries issued by North American, European and Asian peers, subsequently, for periods of 1,2, and 4 days. Furthermore, we determine the pairwise intersection between the query sets and the intersection of all three sets. The cardinalities of the sets for the three time periods are shown in Table 3. We note that the cardinality of the intersection between the query sets of North American and European peers is about 2.8% of the cardinality of the North American set and the European set for a single day. Even for a 4-day period, the cardinality of the intersection is not larger than 6%. The relative cardinality of the intersection of the query sets from all three continents is between 0.001% and 0.02% for all geographical regions and periods. We conclude from Table 3.8 that peers from different geographical regions issue different queries. Nevertheless, there is a small intersection that should be considered in an accurate workload model.

As a consequence of Figure 3.20 and Table 3.8 we employ the following methodology for computation of representative query popularity distributions. To account for geographical correlations, we divide the queries for each day into seven sets, i.e. one set for queries that are issued only from a single geographical region, three sets for queries that are issued by peers from two geographical regions (one for each pair), and one set of queries that are issued by peers from all three regions. We rank the queries by their frequency for each day and each of the seven subsets. To consider the hot set drift, we calculate the average frequency for a query with rank i for all days. Note, that according to Figure 3.20 the query with rank i on day n in general is different from the query with rank i on day $n + 1$. Thus, ranking queries separately for each day preserves the hot-set drift and we obtain an average distribution of query popularity for a single day.

The small number of distinct queries issued by all Asian peers per day (on average 153) does not provide a representative and stochastic meaningful sample. Thus, we concentrate the characterization of the query popularity on North American and European peers in the remainder of this section. Figure 3.21 plots the probability mass function (pmf) of the popularity distributions for the class of queries issued only by North American peers, the class of queries issued only by European peers, and the class of queries issued by both North American and European peers. On a log-log scale, the curves are nearly linear, indicating that the query popularity per day follows a Zipf-like distribution. Note that the skew in the Zipf-like distribution (i.e. the slope of the line) is somewhat different for each region; in fact, the fitted Zipf-like distribution has parameter $\alpha_{NA} = 0.386$ for queries issued only in North America, and $\alpha_E = 0.223$ for queries issued only in Europe. We also note that, similar to [KWX01] and [Sri01], when we compute the popularity distribution for the aggregate set of queries over multiple days from our measurement trace, we observe a flattened head in the distribution, since

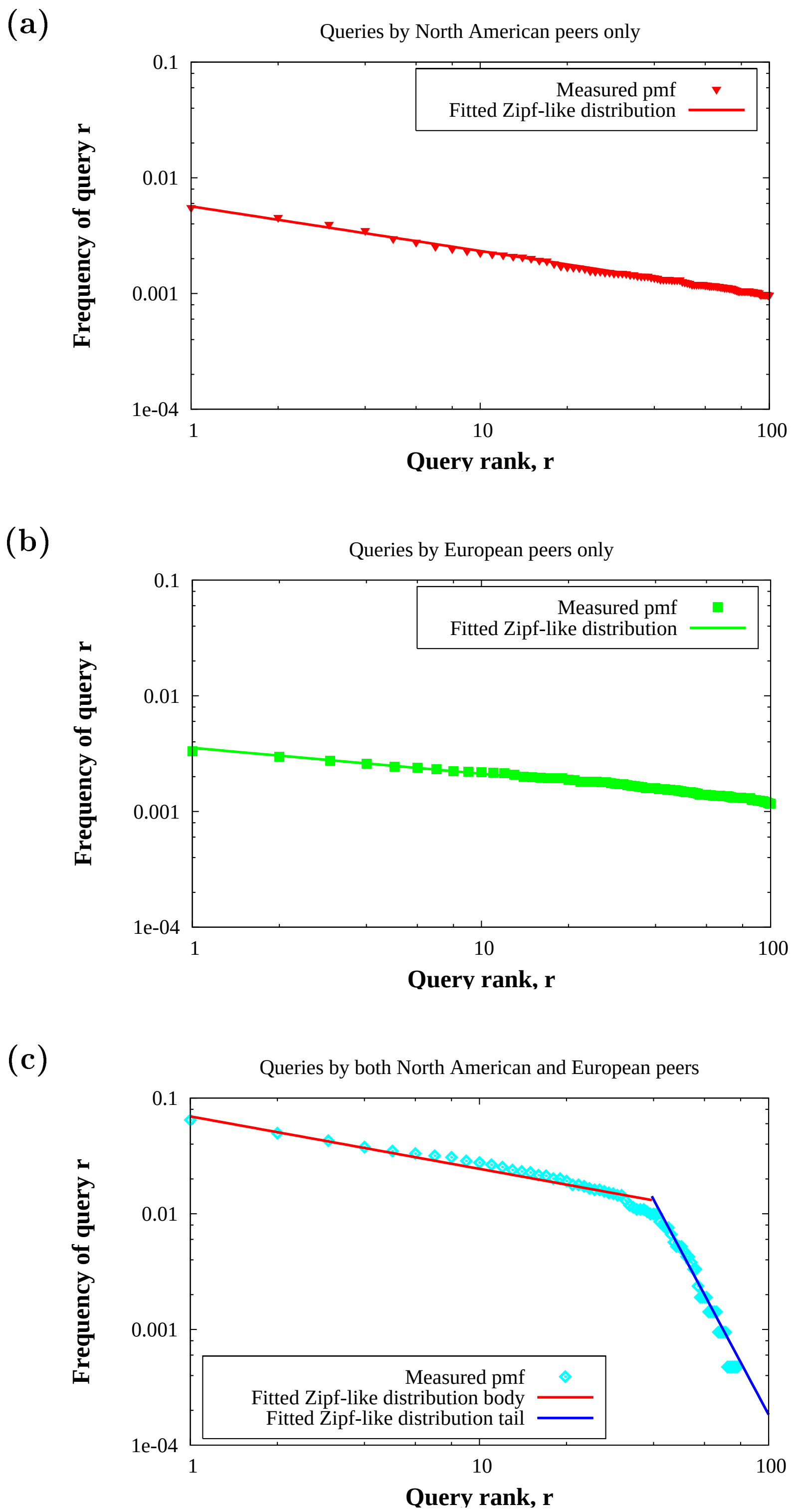


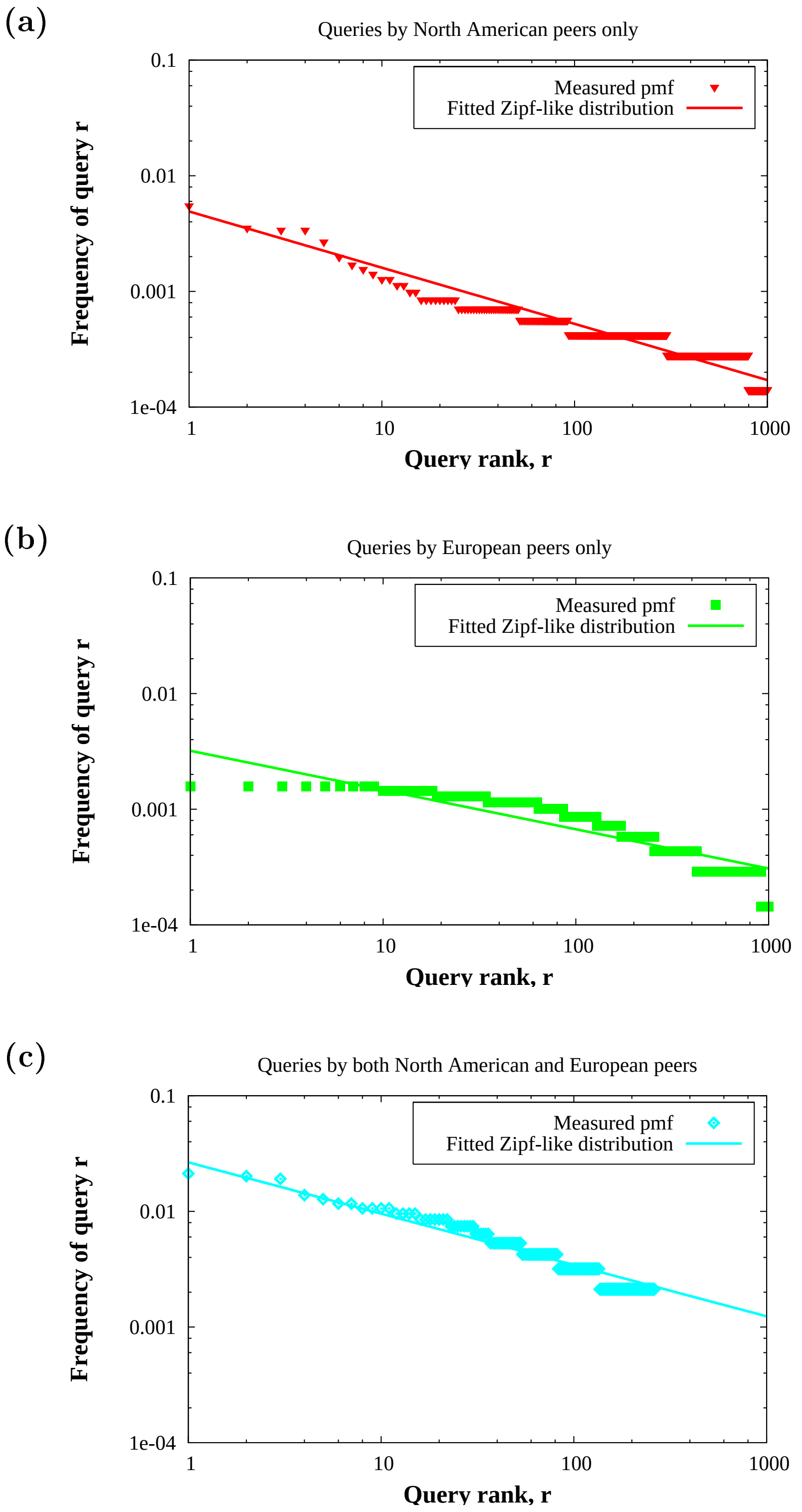
Figure 3.21: Distribution of average 1-day query popularity

there are multiple queries in the aggregate set that were accessed with similar frequency but on different days. This can be seen in the popularity distributions for multiple days periods of 4 and 10 days that are shown in Figure 3.22 and Figure 3.23, respectively. Similarly, the popularity distribution for the queries that are issued by both North American and European peers (shown in Figure 3.21 (c)) has a flattened head and is fit by two different Zipf-like distributions, one for queries ranked 1 to 45 with $\alpha_{I,body} = 0.453$ and the other for queries ranked 46 to 100 with $\alpha_{I,tail} = 4.67$. Furthermore, the values of these Zipf parameters are significantly smaller than those observed in related work [Sri01], due to filtering of automated queries. This fact, again, provides evidence that the filtering of automated client behavior is essential for characterizing user behavior in a peer-to-peer file sharing system. As a consequence of the small Zipf parameters, caching of responses will be more effective for reducing network load in systems that use aggressive automated re-query features than in systems that only issue queries on the users action.

For synthetic workload generation for peers in North America and Europe, the results shown in Table 3.8 and Figure 3.21 can be used as follows: For North American peers, a query is in the set of North American queries with a probability of 0.97, and with probability 0.03 in the intersection set of queries from North American and European peers. Thus, for each query the set is chosen with these probabilities. After that, the query is chosen by a Zipf-like distribution with the parameter determined by Figure 3.21 for the according set. Recall that we use only two geographical regions, North America and Europe in the characterization of the query popularity. However, the methodology can be extended in a straightforward way to more geographical regions.

3.4.7 Generating Synthetic Workloads

In this section we outline how to apply the characteristics derived in Sections 3.4.1 to 3.4.6 to generate P2P file sharing system workloads. Consider a system in steady state with N peers. When a peer finishes a session, it is replaced by a new peer. All peers are modeled as described in Figure 3.24. The evaluation is performed for a given time of day, which is selected before workload generation. The algorithm can be applied for peers from each geographical regions by using the corresponding conditional distributions identified in Sections 3.4.1 to 3.4.6.



(a) Queries by North American peers only

(b) Queries by European peers only

(c) Queries by both North American and European peers

Figure 3.22: Distribution of 4-day query popularity

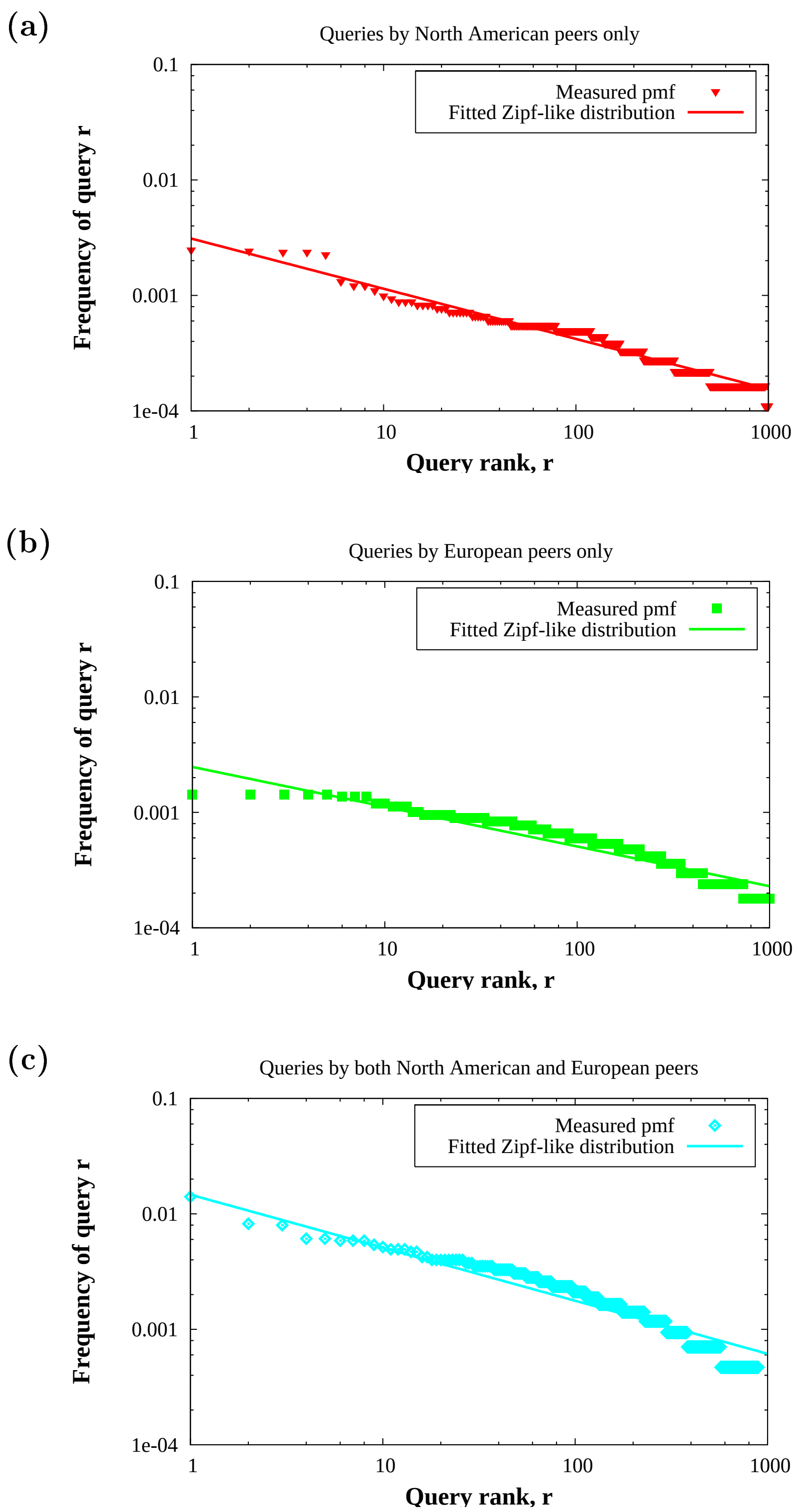


Figure 3.23: Distribution of 10-day query popularity

For each peer:

(1) Select the geographical region, each with probability conditioned on time of day as given by Figure 3.1.

(2) Determine whether the peer is passive or active, each with the probability conditioned on geographical region, as given by Figure 3.4.

(3) *If the peer is passive:*

- (a) Choose the connected session duration conditioned on time of day according to Table 3.3.

(4) *If the peer is active:*

- (a) Determine the number of queries per session conditioned on geographical region according to Table 3.4.
- (b) Determine the time until first query conditioned on the number of queries and time of day according to Table 3.5.
- (c) *For each query:*
    - (i) Determine the query interarrival time conditioned on time of day according to Table 3.6.
    - (ii) Determine the class of the query according to Table 3.8.
    - (iii) Determine the rank of the query according to the distribution of rank for the query class, e.g., according to Figure 3.21 (a) for queries by North American peers only.
- (d) Determine the time after last query conditioned on time of day according to Table 3.7

Figure 3.24: Algorithm for generating a synthetic workload

3.5 Summary

In this chapter, we provided a detailed synthetic workload model for the query behavior of peers in a P2P file sharing system. We conducted a comprehensive measurement study in the Gnutella system in order to capture the query behavior of individual peers. By developing filter rules that distinguish user generated queries from queries generated automatically by the client software, the system-independent user behavior is extracted from the trace. Based on the measured data, we characterized the user-induced query behavior by analyzing various workload measures. These measures include the fraction of connected sessions that are completely passive, the duration of such sessions, and for each active session, the number of queries issued, the query interarrival time, the time until first query, the time after last query, and the query popularity. The characterization captures the key correlations in the observed workload measures as well as correlation with geographical region and time of day, such that the measured behavior can be used in constructing realistic synthetic workloads.

Key new observations include: (1) automated re-queries generated by the client software have a significant impact on most workload measures and thus have to be filtered for characterizing user behavior, (2) number of queries per session and passive session duration are sensitive to geographical region, with peers in Europe issuing on average more queries per session and having longer passive session durations than peers from North America and Asia, (3) the time between the last query in an active session and the end of the session has a much heavier tail than the time between queries, (4) the 100 most popular queries change significantly from one day to the next and, (5) 97% of the queries issued from peers in North America are not issued by peers in Europe, and vice versa.

Considering all correlations between the workload measures, the geographical region, and the time of day, we provided the fitted conditional distributions and the corresponding parameters which accurately reflect the measured data. Furthermore, an algorithm for generating detailed query workload is given which can be directly used for evaluating new P2P system designs.

Chapter 4

Characterization of Web Sites by their Structural Properties

The classification of individual Web sites within a particular top-level domain will be extremely valuable for improving the capabilities of search engines. In fact, the classification of individual Web sites will support personalized ranking, clustering of search results, and search in specialized indexes. To build the basis for a coarse classification of Web sites which does not employ costly content-based classification approaches, this chapter analyzes the structural properties of Web sites. These structural properties reflect the size, the organization, the link structure, and the composition of URLs of Web sites. Contrary to previous work, we analyze entire Web sites instead of individual Web pages. Our measurement study is based upon more than 2,300 Web sites composed of 11 million crawled and 73 million known pages from the top-level domain `.de`. To provide a comprehensive picture of the German Web, we distinguish twelve different categories of Web sites. Building on the presented analysis of structural properties, we can aggregate these twelve categories into the five classes *Brochure*, *Listing*, *Blog*, *Institution*, and *Personal* considering close similar structural properties and similar function at the same time. For these classes of Web sites the structural properties are compared and characterized in terms of probability distributions which well reflect the measured data. By investigating the correlation structure between the structural properties we identify further differences in the composition of Web sites of different classes.

4.1 Previous Results on Characterization of Web Sites

There is a large body of prior work on measurement and characterization of the Web which addresses various aspects. Most of these are related to the measurement of Web traffic in order to characterize client behavior as in [BC98], Web site usage as in [CG04], [ISZ99], or Web performance as in [BC99]. A somewhat outdated survey over papers considering these aspects of the Web can be found in [Pit99]. Since the measurement and characterization presented in this chapter concentrates on structural properties of Web sites we do not elaborate on the traffic related aspects of the Web.

A further research issue for previous Web characterizations is the evolution of Web pages, i.e. the frequency and degree of how Web pages change over time. Several recent papers, e.g. [CGM00], [NCO04], [FMNW03], presented Web measurement studies to investigate the change frequency and the lifespan of individual Web pages across different top-level domains. Cho and Garcia-Molina [CGM00] studied the evolution of more than 500,000 Web pages drawn from 270 US Web sites. They found that about 40% of all Web pages they considered changed within a week and that it took 50 days until half of the considered number of pages had changed. In a complementary study [FMNW03], Fetterly and his co-workers studied the evolution of more than 150 million Web pages across European, US, and Asian top-level domains by crawling them once a week over a period of 11 weeks. They observed a strong relationship between the top-level domain and the frequency of change of a document whereas the relationship between the top-level domain and the degree of change is much weaker. Recently, Ntoulas, Cho, and Olston [NCO04] presented the most extensive study using weekly snapshots of 154 US Web sites over the course of a year. Among other things, they investigated the predictability of the degree of change for individual Web sites and concluded that the ability to predict future degree of content change from past behavior can vary significantly from site to site.

Opposed to [CGM00], [NCO04], [FMNW03], our Web measurement study is not focused on the evolution of the Web but on analyzing structural properties of Web sites. The evolution studies state that pages drawn from Web sites belonging to different domains change at different rates [CGM00], [FMNW03]. We believe and show that Web sites from distinct thematic clusters, which are comparable to Web sites from the top-level domain `.gov` in contrast to Web sites from `.edu`, also differ in their structural properties.

Characterizations of structural properties so far focus on examining the link structure of the Web. The papers [BKM$^+$00], [DKM$^+$02] report large-scale measurement studies of the Web's link structure to gain insight into its composition. Broder and his co-workers [BKM$^+$00] showed that the distributions of the number of incoming and outgoing links of a page, denoted as *indegree* and *outdegree*, follow power-law distributions. Furthermore, the authors provided evidence that the macroscopic structure of the Web comprises three main components: IN, SCC, and OUT. Thereby, SCC is a strongly connected component containing Web pages in which each page can be reached by each other page in the component following hyperlinks. The IN component comprises Web pages from which SCC can be reached via hyperlinks, but which cannot be reached from pages of SCC. Likewise, OUT contains all pages which can be reached from pages in SCC but do not link to SCC. That is, the Web looks like a bow tie. Building on these results, Dill et al. [DKM$^+$02] discovered that both micro- and macroscopic graph structures of the Web possess the bow tie structure, i.e. cohesive sub-regions of the Web display the same characteristics as the Web at large with respect to the link structure. These cohesive sub-regions, introduced as so called *thematically unified clusters*, are for example collections of Web pages sharing a common subject or a random collection of Web sites.

We adopt the idea of considering thematically unified clusters. Though, instead of clustering single pages or building random collections, we identify categories of functional related Web sites. In contrast to [BKM$^+$00] and [DKM$^+$02], we focus on measuring and analyzing structural properties of entire Web sites. The aim of our study lies in figuring out how Web sites are composed and in identifying differences between sites of distinct categories. This enables the coarse classification of Web sites by their structural properties. In contrast, [BKM$^+$00], [DKM$^+$02], aimed at gaining insight in the graph structure of the Web.

First characterizations of entire Web sites include [MS97] and [BRVX04]. Manley and Seltzer analyzed 10 Web server log files representing 3 different categories of Web sites in order to compare the access patterns and growth of these sites [MS97]. They found that depending on the site category, the increase in number of visits of a site correlates with the number of Web users, the number of documents a user visits on the site, the fee structure of the site, the visibility in search engines, or marketing strategy of the Web master. Thus, the authors showed differences in the usage patterns of sites belonging to different categories. In [BRVX04] Bent et al. characterized the workload at 3000 commercial Web sites in terms of cookie usage and cachability of documents. They found that the indiscriminate usage of cookies for all types of documents (even pictures) is very common in the observed set of Web sites. Furthermore, most Web sites do not use the cache-control features of HTTP 1.1. The authors conclude, that the performance could be significantly improved if the Web masters would more carefully consider the performance implications of their Web design.

As [MS97] we identify differences between sites of different categories. However, in contrast to [MS97] and [BRVX04], we do not only consider academic, business, and informational sites, but provide a characterization of a comprehensive set of Web sites which represents all major categories of the Web. Furthermore, as [BRVX04] we analyze the organization of Web sites but do not focus on caching-related measures. Instead, we provide the basis for automatically classifying Web sites into categories based on a number of structural properties.

Recently, Gibson et al. studied the volume and evolution of Web page templates [GPT05]. Template material is common content or formatting that appears on multiple pages of a Web site. They found that 40-50% of the content on the Web constitutes templates. Furthermore, they observed that template usage significantly differs for different categories of Web sites. As [GPT05], we analyze structural properties of entire Web sites and realize the need for the coarse classification of Web sites. In contrast to [GPT05], we identify the major classes of the Web by clustering functionally related categories of Web sites based upon structural properties while still keeping the functional relation.

Closely related to the characterization presented in this chapter is the paper [ACD$^+$03] by Amitay et al. Based on a measurement study of 202 large and popular Web sites in late 2001 they analyzed the link structure and directory organization of sites from eight categories. Without characterizing the sites in detail Amitay et al. report differences in several structural properties for sites from different categories. These structural properties are related to the distribution of pages into the site's directory structure, the external indegree and outdegree, and the internal linkage structure of the site. A decision-rule classifier exploiting the different structure of sites from different categories yields an average classification accuracy of 55% in estimating the site category from the structural properties. This early work reports that the site category correlates to some extent to a number of structural properties. However, Amitay et al. considered static pages only and concentrated on link structure and directory organization. Furthermore, the measured data is heavily biased towards large and popular sites, because the considered sites are chosen according to external indegree and outdegree. In contrast, this chapter provides a detailed characterization of the structural properties of Web sites which examines each individual property and the correlation structure between the properties. This characterization is based on a recent measurement study comprising more than 2,300 Web sites which covers not only linkage related properties but additionally includes structural properties reflecting the size, the organization and the composition of URLs. Furthermore, our study is not restricted to static pages, but also considers dynamically generated documents which make up an increasing fraction of the pages in the Web.

4.2 Measurement Methodology for Structural Properties of Web Sites

4.2.1 Selection of Web Sites

The Web consists of a vast amount of information available on innumerable Web sites. Nevertheless, many of these Web sites are related by the kind of information they present so that they can be assigned to certain categories. The major categories of Web sites are:

Academic: Web sites of universities, research institutions, and universities of applied sciences.

Blog: The group of Web logs.

City: The Web presences of cities.

Education: Web sites of schools and educational institutions like advanced training centers.

Forum: Online forums, discussion groups, and chat portals.

Foundation: Web sites of foundations and non-profit organizations.

Government: The Web presence of departments and federal states.

Information: News and information portals like the Web sites of newspapers or broadcasters.

Medical: Web sites of hospitals, nursing homes, etc.

Private: The homepages of individuals or small groups.

Shop: Online shops and auction portals listing products usually for sale, but also dating services specifying the profiles of numerous users.

SME: The Web presence of small and medium-sized enterprises.

We intentionally omit the categories comprising porn and spam sites. Considering the tremendous size of the Web and the various genres and modes of content that occur on it, this set of categories cannot be complete. However, our experience in analyzing the Web and an inspection of the categories listed in major Web directories (e.g. [Net05], [Yah05]) let us conclude that these are the major categories most of the Web sites can be assigned to.

In order to analyze the structural properties of Web sites in dependence of their category, we select several Web sites from each category by randomly choosing their corresponding URL from different Web directories. Most of the URLs are obtained from the Open Directory Project [Net05] by selecting an appropriate category. As representation for the Shop category, we gathered the URLs of trusted online shops listed in [Tru05] while the Web sites belonging to the category of small and medium-sized enterprises are taken from the Hoppenstedt directory [Hop05].

Our measurements are strictly focused on the German part of the Web to avoid noisy data due to mixing Web sites from different countries whose structure and organization might be influenced by national distinctions. Therefore, we only obtain URLs from the top level domain `.de`. In addition to this, we select only URLs pointing at the entry page of a Web site, i.e. URLs without a subdomain or a path leading to a page in a subdirectory. This approach guarantees that the data for each Web site is collected beginning with the entry page of a Web site. After selecting the set of URLs for each category, every Web site is manually examined to assure that it really belongs to the assigned category.

4.2.2 Collecting the Data

For collecting data from the selected Web sites, we employ a search engine software system developed by our group. Our search engine is capable of crawling and indexing about 50,000 pages per hour on a Linux dual-processor PC server with 3.0 GHz Intel Pentium IV Xeon processors and 6 GB RAM. The data was collected in August 2005. The crawl is seeded from the sets of URLs of the Web sites belonging to the identified categories. These sets are disjoint so that every Web site belongs to exactly one category. We define a *Web site* or *site* as the set of Web pages which belong to the same domain, e.g. `uni-dortmund.de`. Thus, according to our definition the pages located in a subdomain of a Web site, e.g. `cs.uni-dortmund.de`, are also considered as belonging to this Web site.

Starting with crawling the entry page of a Web site, the content of the page is parsed to extract links to further pages. Afterwards, these linked pages are crawled and parsed and so on. Our crawler scans each single Web site in a breadth-first-search manner following only internal links, i.e. links pointing at a page within the same domain, while external links are counted for later analysis but discarded afterwards. By crawling the Web sites in this way, we assure on the one hand that only pages from the pre-selected Web sites are downloaded and considered for our studies. On the other hand, we are able to determine the level of a page, i.e. the minimum number of clicks it is apart from the entry page. The source code of the page is stored in a repository together with some additional information about the page. This information consists of the URL, a unique document ID, the size of the document, the crawling date, and a status code, indicating whether or not the document was crawled correctly.

To reduce the traffic placed on the servers hosting the selected Web sites, we crawled at most 20,000 pages per Web site or at most 2 GBytes of data. These boundaries allow collecting data in a sufficient way as most Web sites comprise less than 20,000 pages and less than 2 GBytes of textual content. While these boundaries limit the characterization of the Web sites in terms of number of crawled pages, we use additional data from further discovered but not crawled pages of these Web sites as described in the following section. In addition to this restriction, our crawler obeys the robots exclusion protocol and the netiquette by keeping a timeout of at least two seconds between two successive requests to the same server. The crawl is completed when no further pages belonging to the

Category	**#Web sites**	**#Crawled pages**	**#Known pages**
SME	607	619,256	2,328,523
Shop	376	3,395,151	31,970,563
Private	279	275,217	626,284
Blog	224	766,072	2,268,926
Academic	158	2,233,615	12,468,358
Education	149	126,381	752,420
City	134	1,550,264	8,143,157
Medical	124	183,169	1,102,234
Information	81	990,887	6,725,939
Foundation	80	153,272	1,301,321
Forum	79	800,177	4,651,418
Government	28	467,825	1,440,935
Total	2,319	11,561,286	73,780,078

Table 4.1: Measurement statistics of categories

pre-selected Web sites can be retrieved obeying the restrictions described before.

Table 4.1 shows the number of Web sites per category with at least 100 crawled pages. Furthermore, it summarizes how many pages have been overall crawled per category and how many pages are known. The number of known pages includes the number of crawled pages plus pages belonging to the Web sites within the category, which have been discovered (by examining the URLs extracted from the crawled pages) but have not been downloaded. We denote the number of these pages as *known pages*. All in all we analyze the structural properties of 2,319 Web sites from twelve distinct categories. The analysis is based upon 11,561,286 crawled and 73,780,078 known pages.

4.2.3 Representativeness of Measured Web Sites

Note that the measurement study is biased by a number of limitations to the choice of the considered Web sites. First, we concentrate on the German part of the Web, not at least due to the inability of the author to precisely categorize foreign language Web sites. Thus, the results obtained in this chapter are valid for German Web sites only. However, the methodology developed for the measurement and characterization can be employed for the Web of other countries as well.

Second, as stated above, the Web sites are sampled from existing directories which may already be biased by some unknown selection criteria. While the Hoppenstedt directory [Hop05] for small and medium enterprises (SME) provides a roughly complete picture of the SME sites in Germany, the directories in [Net05] and [Tru05] comprise only a subset of all Web sites. Nevertheless, since the goal of this chapter is to identify qualitative differences in the structural properties of Web sites from different categories and the number of sample Web sites from each category is sufficiently large, we argue that the measured Web sites present a representative picture of their particular category for the Germany Web.

Third, we only analyze Web sites from which at least 100 pages could be crawled correctly to base our analysis on statistical significant sample sets of pages within each Web site. While this also biases the selection of Web sites for our analysis, the sites with less than 100 pages missed by our characterization reflects "less important" Web sites with respect to the number of pages. Summarizing, due to the reasons stated above we strongly believe that our measurement study is representative for the German Web.

4.2.4 Structural Properties of Web Sites

Measures for capturing structural properties of Web sites should reflect their size, their organization, their link structure as well as the general composition of the URLs of their pages. Therefore, we identify characteristic measures for each type of structural properties.

The size of a Web site is captured by two measures. On the one hand we determine the number of known pages per Web site reflecting the size in terms of page-count and on the other hand we calculate the average document size describing the size in terms of the amount of available data. Note that only documents comprising textual content like HTML, PDF and Text documents are considered. That is, pictures, sound and video files are not included in the computation of the average document size.

The organization of a Web site is spotted by counting the number of distinct subdomains per Web site, analyzing the fraction of document types, and by detecting the average level of its pages. We compute the number of subdomains by adding up the number of different domain parts of the URLs of the corresponding Web site. The domain part of an URL is given by the text between the protocol specification `HTTP://` and the first slash after the top-level domain. The document type is determined by the file extension within the URL. We parsed the URLs for many of the most common file extensions including `.html`, `.xml`, `.txt`, `.pdf`, `.ps`, `.php`, `.asp`, `.jsp`, `.pl`, but concentrate our analysis on the fraction of HTML documents. The average level of a Web site is determined by computing the level of every page, i.e. by counting the minimum number of links that have to be followed beginning from the entry page in order to reach this page.

Obviously, the link structure provides another source for further insight into the structural properties of a Web site. We consider the link structure by calculating the average internal and external outdegree. The *outdegree* of a page is defined as the number of links within a page pointing at other pages belonging to the same Web site, i.e. internal links, or pointing at pages on other Web sites, i.e. external links. The sum of the internal and the external outdegree is the overall outdegree. Duplicate links within one page, i.e. several links pointing at the same other page, are counted only once. Note that we do not consider *indegrees*, i.e. number of links pointing to a specific page or site, because this measure heavily depends on the measured snapshot of the Web. That is, a correct value for the indegree can only be calculated if all pages of the entire Web have been crawled. Since this is impossible, the indegree is always underestimated. Furthermore, the external indegree for a given site is generally not controlled by the Web master of this site, but by other Web masters. Thus, this measure does not reflect a structural property of the site. Additionally, the indegree of a page depends to some extent on the age of the page. I.e. the newer a page is, the smaller is the probability that other pages link to it.

Further properties describing the general composition of the URLs of a Web site can be directly derived from the URLs. We determine the average length of the URL path, i.e. the part of the URL behind the top level domain. This measure provides some indication about the organization of the site's pages in the Web server's file system. For example, different directory and file names result in different path lengths. Furthermore, we count the number of slashes and digits within the path. The number of slashes gives insight into the directory structure of the Web server's file system. The more slashes the URLs of a site contain,

Measure	Type
Number of known pages	Size
Avg. level	Organization
Number of subdomains	Organization
Avg. number of slashes in URL path	URL
Avg. number of digits in URL path	URL
Avg. length of URL path	URL
Fraction of HTML documents	Organization

Table 4.2: Measures derived from known pages

Measure	Type
Avg. internal outdegree	Link structure
Avg. external outdegree	Link structure
Avg. document size	Size

Table 4.3: Measures derived from crawled pages

the deeper is the file system where the pages are stored. However, a significant fraction of Web sites use dynamically generated Web pages, for example to present customized pages to individual users. These sites often use session IDs or product IDs in the URLs. Since these IDs in general contain many digits, we account for this structural property by measuring the average number of digits in the URLs of a site.

Tables 4.2 and 4.3 summarize all measures for the different types of properties and state whether a measure can be derived from all known pages or just from the crawled pages of a Web site. Since most of the measures can be derived from known pages, the available amount of data for the analysis grows rapidly.

Although our analysis is focused on the German part of the Web, we strongly believe that our methodology is applicable to other parts of the Web, too. This holds especially for Web sites within the top-level domain of other industrialized countries. Since the identified structural properties are independent of a page's content, in particular the used language, they can be determined easily for the Web sites representing the major categories of the Web. These Web sites can be obtained from trusted Web directories for the considered top-level domain. As a consequence, applying our measurement methodology to any top-level domain gains insight into the structural properties of Web sites from distinct categories and forms the basis for their coarse classification.

4.2.5 Outlier Analysis

As discussed in Section 2.1 outlier analysis provides a good starting point for identifying potential measurement errors or anomalies in the measured data. This is especially important for Web measurements, because Web sites are maintained by individuals with very different technical skills, intentions, and Web publishing tools. Furthermore, while the programming of individual Web pages is to a small extent regulated by the HTML standard, the organization of Web sites is completely up to the Web master. E.g. some spamming Web sites deliver different Web pages to Web crawlers than to conventional browsers like the Microsoft Internet Explorer. This is intentionally done to improve ranking of the site in search engines like Google [Goo05], Yahoo! [Yah05], or MSN Search [Mic05]. Thus, Web sites are extremely diverse and hard to measure, because the measurement software has to account for various programming styles of Web sites.

The outlier analysis reveals several Web sites with extreme values for the observed measures, partly because of measurement errors, partly because of spamming, and partly because of just unusual behavior of specific sites. By carefully inspecting those outliers, we identify five sites with average number of slashes of more than 30 which are hosted at a specific Internet service provider (ISP). This ISP tries to improve ranking in search engines by returning faked HTML pages with recursive links for each request to the crawler. Human users using a browser see "ordinary" Web pages. This spamming technique is known as *cloaking* [GGM05]. We discard those sites from our analysis, since the measured structure does not correspond to the real structure of the site.

A further problem occurs with two sites which use dynamic URLs containing hidden session IDs. These sites do not use ordinary session IDs which appear as parameters in the URL, but compile the sessions IDs into the filename of the Web document. For these sites we measured very large values for the number of known pages, although the number of crawled pages is rather small. This behavior is caused by the inability of the crawler to identify this special style of using session IDs, i.e. constitutes a measurement error. Note that the crawler generally supports session IDs as used by many sites. These two sites are also discarded.

Analyzing the number of subdomains of each site, we observe three sites with more than 500 subdomains. These sites (two online shops and one site hosting numerous Web logs) have an own subdomain for each product or Web log, respectively. Although very unusual, this behavior is no measurement error. Thus, we keep these sites for later characterization. For all remaining structural properties we verified that outliers are not caused by measurement errors, but reflect the real structure of the Web sites.

4.3 Characterization of Web Site Structure

In this section, we provide insight into the structural properties of Web sites belonging to different categories and provide evidence that sites from different categories differ significantly. Besides gaining deeper understanding of how Web sites are composed, the goal of the characterization is to provide the basis for an automated classification of Web sites. The advantage of such a structural property-based classification is that it has much less processing demands than the content-based approaches of previous classifiers, e.g. [EKS02], [KS04], and [GTL$^+$02]. To keep this thesis self-contained, Section 4.5 sketches such an automated classifier for Web sites. However, the focus of this thesis lies on the measurement and characterization of Internet applications.

We will show that the Web site categories identified in Section 4.2.1 can be clustered into classes by considering both structural properties and functional relation. Subsequently, we compare the structural properties of Web sites from each class and characterize each property providing model distributions which reflect the measured data. Finally, the correlation structure between the structural properties is analyzed. Note that although there are several correlations between the structural properties, the model distributions are given for each measure disregarding these correlations. The reason for this is that easily applicable classifiers as e.g. the naïve Bayesian classifier [DHS01] make use of the independence assumption of discriminators. However, it is shown in [DP97] that these type of classifiers perform well in practice even when the independence assumption is violated.

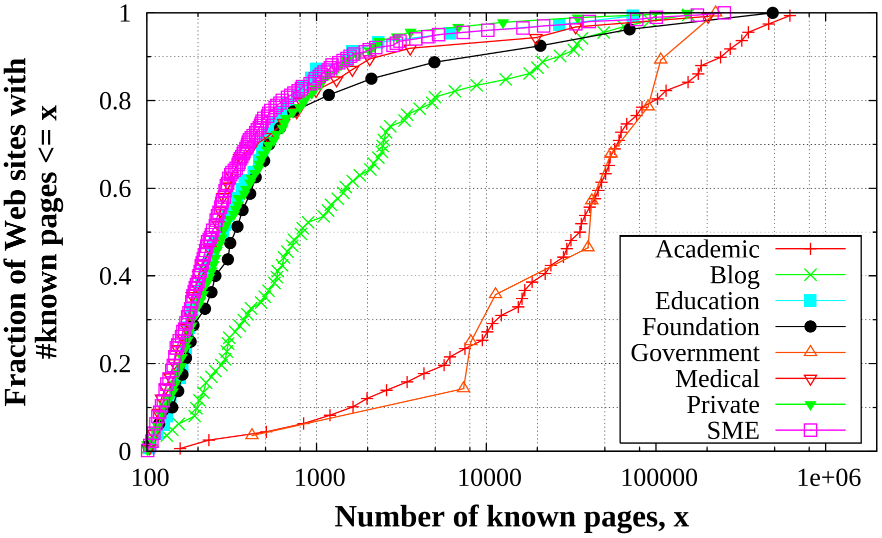


Figure 4.1: Distribution of number of known pages for various categories

4.3.1 Clustering Categories into Classes

We analyze the structural properties of the Web sites assigned to the categories identified in Section 4.2 by plotting the cumulative distribution of each measured parameter. In fact, there are significant differences in the distributions of the structural measures between the categories. However, we observe that particular subsets of the categories have similar distributions in many measures, i.e. some groups of categories have similar structural properties. Consider for example the distribution of the number of known pages per site shown in Figure 4.1. Note that each curve is plotted with specific symbols for better readability. However, the number of symbols for each curve does not represent the number of measured data points. We observe that the sites from categories Academic and Government have similar number of known pages and on average much more pages than the sites of other categories. A second subset containing sites from Education, Foundation, Medical, Private, and SME has also similar distributions of the number of known pages. These sites are on average smaller than sites from the Blog category. Note that for ease of readability not all 12 categories are included in the figure.

As second example for similar distributions in different subsets of categories,

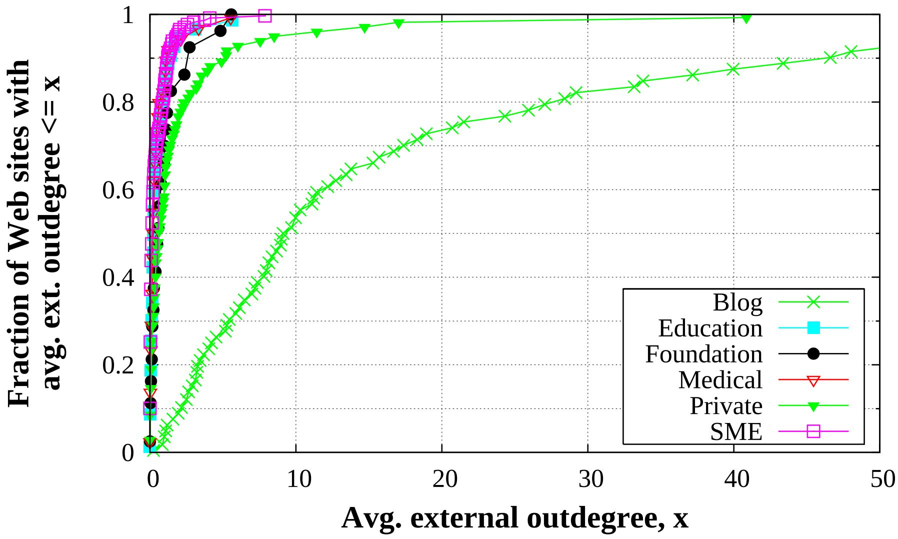


Figure 4.2: Distribution of average external outdegree for various categories

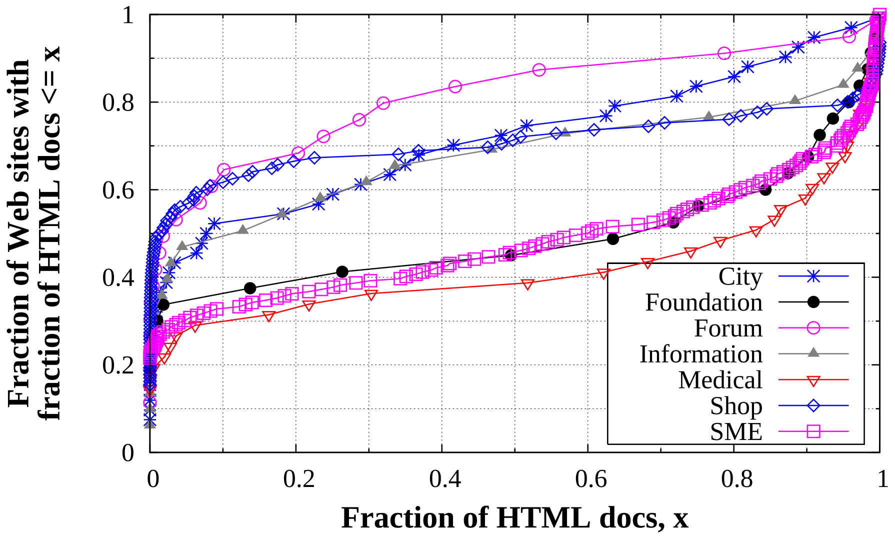


Figure 4.3: Distribution of fraction of HTML documents for various categories

Figure 4.2 plots average external outdegree for various categories. Again the sites from Education, Foundation, Medical, and SME have similar distributions. Sites from the category Private differ from this subset by having on average more external links. As for the number of known pages, Web logs are very different to the other categories. Sites from this category have many links to pages of other sites.

Another subset of similar categories contains City, Forum, Information and Shop, which can be seen in Figure 4.3. The figure plots the distribution of the fraction of HTML documents within a site. We observe that there are similar distributions of this measure for the categories City, Forum, Information, and Shop as well as for the categories Foundation, Medical and SME.

Similar trends can be observed for all of the considered structural properties. Thus, categories which have similar structural properties can be clustered into classes. The allocation of categories to classes, as identified by graphical analysis of the structural properties, is given in Table 4.4. We observe that the categories of a specific class have not only structural similarities but are also functional related:

Brochure: The sites of the categories Education, Medical, Foundation, and SME have in common that their main task is to portray an organization. Thus, we denote this class as *Brochure*.

Listing: The second class denoted as *Listing* consists of the categories City, Forum, Information, and Shop. The common nature of the sites in these categories is that they list large numbers of items, i.e. products in the case of online shops, news in the case of information portals, articles in the case of forums, and administrative information for citizens or local news in the case of cities.

Blog: Web logs in general also provide listings of news entries or photos and could therefore be put into the Listing class. However, since their structural properties differ significantly from the categories in the class Listing, as can be seen in Figures 4.1 and 4.2, these sites build an own class denoted as *Blog*.

Institution: The categories Academic and Government are grouped into the class *Institution* because the sites from both categories reflect non-profit organizations providing high-quality information.

Class	**#Sites**	**#Crawled pages**	**#Known pages**	**Categories**
Brochure	960	1,082,076	5,484,498	Education, Medical, Foundation, SME
Listing	670	6,736,479	51,491,077	Shop, City, Information, Forum
Blog	224	766,072	2,268,926	Web logs
Institution	186	2,701,440	13,909,293	Academic, Government
Personal	279	275,217	626,284	Private

Table 4.4: Allocation of categories to classes

Personal: The fifth class *Personal* consists of sites from the category Private.

In addition to the clustering of categories into classes, Table 4.4 provides statistics about the number of sites, number of crawled pages and number of known pages for each class. It shows that each class consists of at least 186 sites and more than 750,000 crawled pages, thus building a statistically significant subset of the considered classes in the German web.

4.3.2 Measures Reflecting Size of Web sites

Number of known pages

The first measure that distinguishes sites of different classes is their size in terms of number of Web pages. Obviously, large organizations like universities and governmental institutions have in general much larger Web sites than private persons or small enterprises. This intuition is underlined by Figure 4.4 which shows that more than 88% of the Web sites from class Institution have at least 3,000 pages. For Web logs this fraction is only about 25%, and for sites from class Brochure the fraction is about 9%. The class Personal is similar to class Brochure and the class Listing is similar to class Institution in this measure. We conclude that number of known pages of a site significantly differs for different site classes and, thus, provides discriminative power for the automated classification.

According to Section 2.3 we identify the probability distribution function which best models the number of known pages for each class of Web sites by

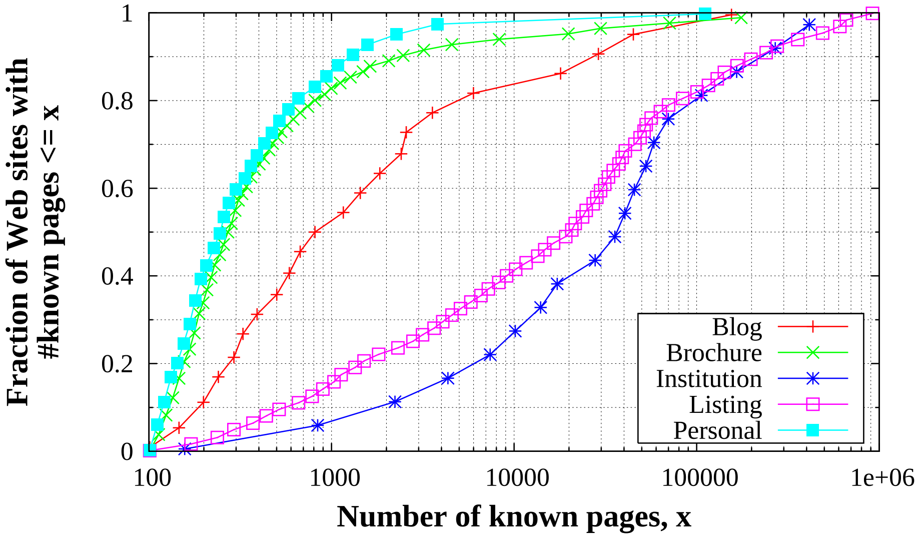


Figure 4.4: Distribution of number of known pages

least-squares regression. Note that we employ the CDF instead of the pdf for fitting, because the CDF is not biased by discretization. The distribution functions and parameters can be directly used for automated classifiers exploiting structural properties of Web sites.

Table 4.5 shows model distributions and the corresponding parameters for number of known pages for each class of Web sites. We observe that number of known pages follows a Weibull distribution for the classes Institution and Listing, whereas this measure follows a lognormal distribution for the classes Blog, Brochure, and Personal. Furthermore, the small values of the error-term Δ given by the root-mean-square of residuals [KK62] indicate a close fit of the distribution functions to the measured data. Recall that we did not consider Web sites from which less than 100 pages could be crawled. Thus, for fitting the probability distributions to the measured data we subtract 100 from the measured values of number of known pages in order to let the distribution start at zero.

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Lognormal	$\mu = 6.782$; $\sigma = 2.093$	0.02534
Brochure	Lognormal	μ= 5.154; $\sigma = 1.565$	0.02372
Institution	Weibull	$\alpha = 0.6948$; $\lambda = 0.0005217$	0.02187
Listing	Weibull	$\alpha = 0.4920$; $\lambda = 0.005870$	0.01829
Personal	Lognormal	μ= 4.962; $\sigma = 1.590$	0.009957

Table 4.5: Model distributions and parameters for number of known pages

Average document size

The average size of the documents of a site provides insight into the amount of hosted content. Figure 4.5 shows that 60% of the sites from class Personal have an average document size of less than 10 kBytes whereas 80% of the sites from class Blog have an average document size between 10 and 35 kBytes. Sites from classes Listing and Institution have similar average document sizes, which are larger than those of the classes Personal and Blog. 40% or 50% of the sites from these classes respectively have average document sizes of more than 35 kBytes. Average file sizes of sites from class Brochure are less bound to a fixed range but are more spread over the range of observed document sizes. I.e. there is a significant fraction of sites from this class which have small file sizes like sites from Personal or Blog and another significant fraction has large file sizes like sites from Listing and Institution. Furthermore, there is a large fraction of about 20% of sites from Brochure which have on average very large files of more than 70 kBytes. Note that only 10% of sites from Listing and Institution have such large files. We conclude from Figure 4.5 that average document size differs significantly between the different classes.

We provide model distributions and the corresponding parameters which match the measured distributions of average document size in Table 4.6. We observe that average document size follows a lognormal distribution for all classes. Nevertheless, there are significant differences in the parameters of the distributions between the individual classes, reflecting the discriminative power of this structural property.

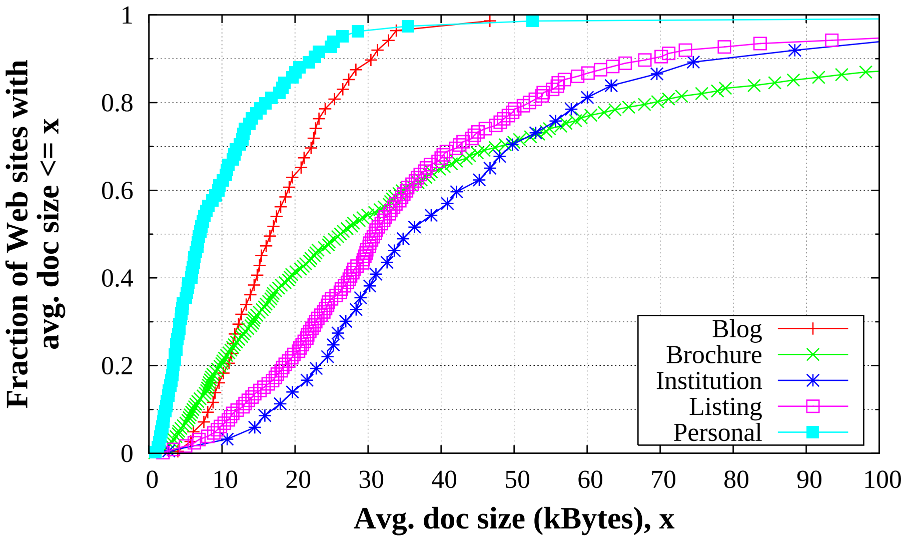


Figure 4.5: Distribution of average document size

Class of Web site	Fitted distribution	Matched parameters	Error term Δ
Blog	Lognormal	$\mu = 9.725$; $\sigma = 0.5094$	0.01529
Brochure	Lognormal	$\mu = 10.19$; $\sigma = 1.189$	0.00532
Institution	Lognormal	$\mu = 10.53$; $\sigma = 0.5877$	0.01325
Listing	Lognormal	$\mu = 10.36$; $\sigma = 0.6068$	0.01463
Personal	Lognormal	$\mu = 8.893$; $\sigma = 0.9151$	0.01462

Table 4.6: Model distributions and parameters for average document size

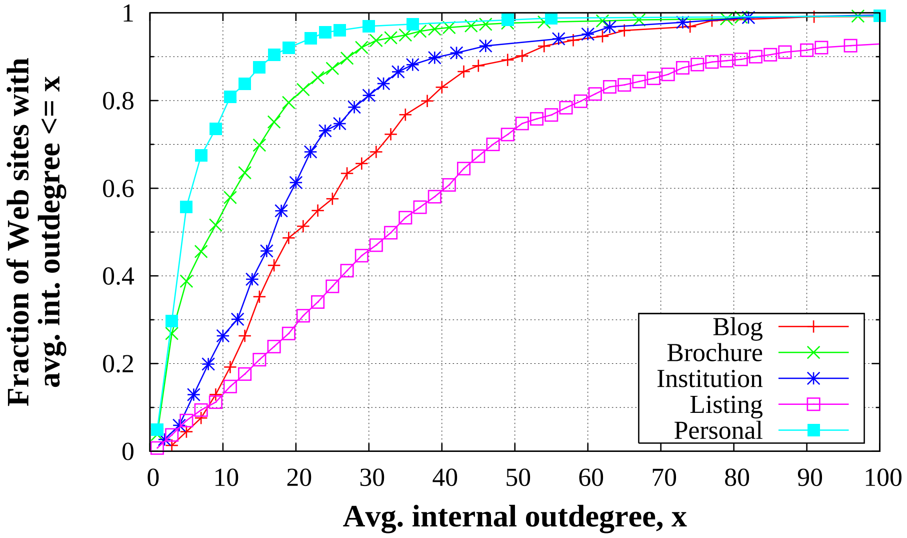


Figure 4.6: Distribution of average internal outdegree

4.3.3 Measures Reflecting Link Structure

Average internal outdegree

Further significant differences between the structural properties of the considered classes can be observed in average internal outdegree depicted in Figure 4.6. Sites of class Listing have on average the strongest internal navigational structure, i.e. provide many links to other pages of the same site. For example about 40% of the sites from this class contain on average more than 40 internal links per page. In contrast, sites of class Personal have on average only few internal links per page. I.e. only about 8% of these sites have on average more than 20 internal links per page. We conclude from Figure 4.6 that average internal outdegree provides significant discriminative power to distinguish between different classes.

Again, the fitted model distributions for each site class are provided in Table 4.7. The table shows that average internal outdegree follows a Weibull distribution for the classes Brochure and Listing, whereas it follows a lognormal distribution for the other classes. The small values for Δ indicate a good match between measured and modeled distribution.

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Lognormal	$\mu = 2.997$; $\sigma = 0.7509$	0.01875
Brochure	Weibull	$\alpha = 0.9661$; $\lambda = 0.09293$	0.01635
Institution	Lognormal	$\mu = 2.784$; $\sigma = 0.6921$	0.01977
Listing	Weibull	$\alpha = 1.341$; $\lambda = 0.006558$	0.01937
Personal	Lognormal	$\mu = 1.545$; $\sigma = 0.9913$	0.01142

Table 4.7: Model distributions and parameters for average internal outdegree

Average external outdegree

The average external outdegree plotted in Figure 4.7 gives additional insight into the composition of Web sites. The figure shows that Web logs have a very dense external linkage structure compared to sites from other classes. This large number of external links is clearly caused by the purpose of most Blog sites to comment events, Web pages, and news articles which are therefore linked. In particular, about 45% of the Blog sites have on average more than 10 links to external pages. In contrast, only about 10% of the sites from class Listing have such average number of external links per page. The smallest average external outdegree are found in sites from Brochure. This can be explained by the fact that especially commercial sites in general have no interest to direct the user to external sites, in particular to competitors. Thus, there are on average relatively few links to other sites.

For the average external outdegree Table 4.8 shows that this structural property follows a lognormal distribution for class Blog instead of a Weibull distribution for the other classes. The observation of this different behavior of the sites from class Blog compared to the other classes corresponds to the same observation made in Figure 4.7. Thus, we conclude that the distributions of average external outdegree for different classes not only differ in terms of distribution parameters but furthermore in terms of the distribution function.

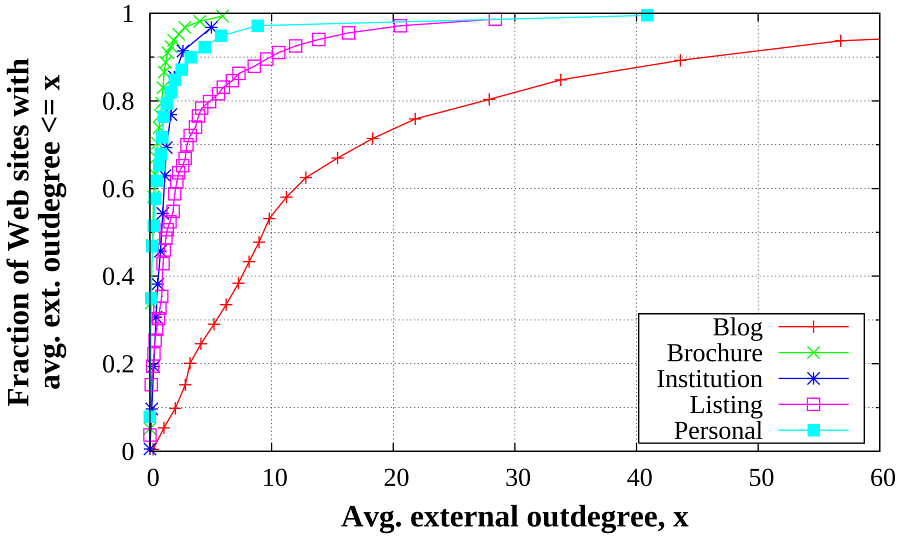


Figure 4.7: Distribution of average external outdegree

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Lognormal	$\mu = 2.228$; $\sigma = 1.172$	0.02103
Brochure	Weibull	$\alpha = 0.6213$; $\sigma = 1.744$	0.01326
Institution	Weibull	$\alpha = 1.166$; $\sigma = 0.7773$	0.01958
Listing	Weibull	$\alpha = 0.6822$; $\sigma = 0.5448$	0.02022
Personal	Weibull	$\alpha = 0.4872$; $\sigma = 1.277$	0.01312

Table 4.8: Model distributions and parameters for average external outdegree

4.3.4 Measures Reflecting Organization of Web sites

Number of subdomains

An important measure for characterizing the organization of Web sites is the number of subdomains forming the site. Figure 4.8 plots the distribution of this measure for all classes. We observe that almost all sites from the classes Brochure, Blog, and Personal have less than 10 subdomains whereas about 5% of the sites from class Listing have more than 10 subdomains. In contrast, sites from class Institution which often have individual subdomains for each department or organizational unit, consist of much more subdomains. Here only 40% of the sites have less than 10 subdomains and more than 10% have more than 200 subdomains. Since the majority of sites from classes Blog, Brochure, Personal, and Listing have only a single subdomain, modeling this measure by a continuous distribution function is inappropriate. However, the observed distribution of number of subdomains for class Institution takes sufficient values to approximate it by a continuous distribution function. Thus, number of subdomains for the latter class can be characterized by a Weibull distribution as stated in Table 4.9. For the other classes the table provides simple discrete distributions which adequately reflect the measured data. We observe from Figure 4.8 and Table 4.9 that only 8% of the private homepages and 7% of the sites from class Brochure have more than a single domain. Furthermore, no site of these classes consists of more than 10 subdomains.

Average level

The average level of a Web site indicates how fast, with respect to number of clicks, information can be found on a Web site. With a small average level, the internal link structure is rather flat, i.e. users need only few clicks from the entry page to a specific page. In contrast, sites with a large average level have a more hierarchic structure. Therefore, users need more clicks to follow a link path specifying the desired information. University sites are a good example of such strongly-structured sites. From the entry page a student looking for lecture material first selects the appropriate department. From the department page the student selects the working group of the lecturer, then he selects the lecturer. From the lecturer's page he gets to the specific lecture which contains the lecture

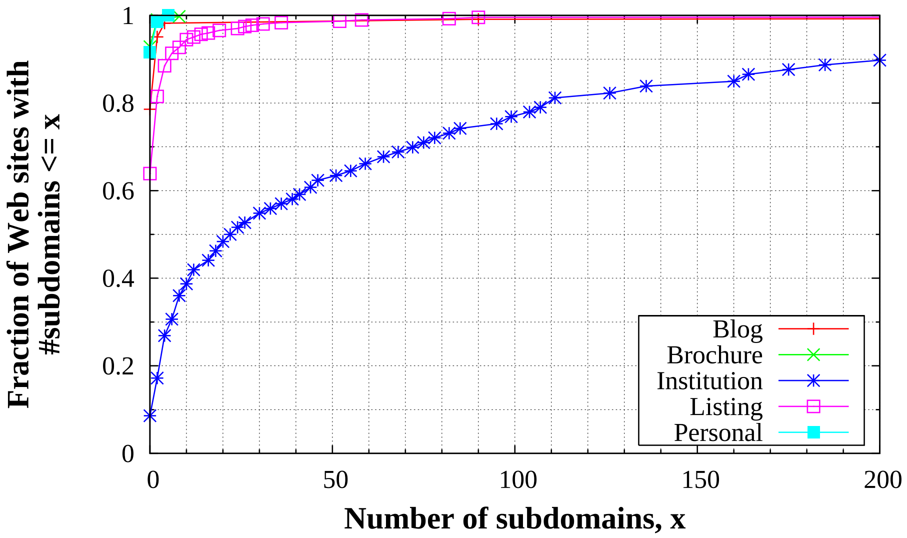


Figure 4.8: Distribution of number of subdomains

Class of Web site	Fitted distribution	Matched parameters	Error term Δ
Blog	Discrete	$P\{X = 0\} = 0.79$	
		$P\{1 \leq X \leq 10\} = 0.20$	
		$P\{X > 10\} = 0.01$	
Brochure	Discrete	$P\{X = 0\} = 0.93$	
		$P\{1 \leq x \leq 10\} = 0.07$	
		$P\{X > 10\} = 0.00$	
Institution	Weibull	$\alpha = 0.5343$; $\lambda = 0.1295$	0.02084
Listing	Discrete	$P\{X = 0\} = 0.64$	
		$P\{1 \leq X \leq 10\} = 0.31$	
		$P\{X > 10\} = 0.05$	
Personal	Discrete	$P\{X = 0\} = 0.92$	
		$P\{1 \leq X \leq 10\} = 0.08$	
		$P\{X > 10\} = 0.00$	

Table 4.9: Model distributions and parameters for number of subdomains

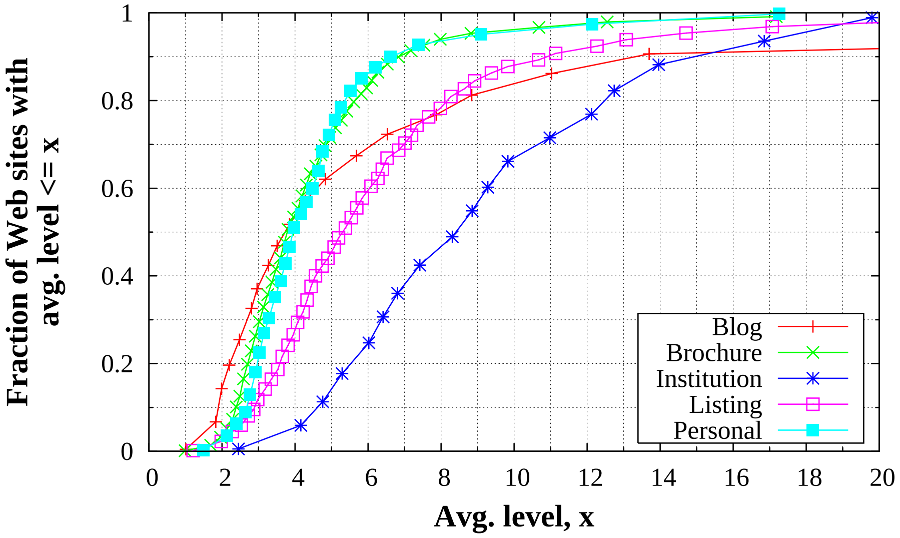


Figure 4.9: Distribution of average level

material. The distribution of average level for each site class is presented in Figure 4.9. Obviously, sites from class institution have in general a larger average level than sites from other classes. Figure 4.9 underlines the attentive reader's presumption that average level correlates with number of known pages, in that classes with large number of known pages (recall Figure 4.4) have also a larger average level. We analyze this and other correlations in detail in Section 4.4. We observe from Figure 4.9 that more than 20% of the sites from classes Listing and Blog have an average level of more than 8. In contrast, the fraction of sites from classes Personal and Brochure with this average level is only 5%. In fact the distributions for classes Personal and Brochure are very similar in this measure.

Table 4.10 provides the model distributions and parameters for average level and shows that this measure can be well modeled by the lognormal distribution for all classes.

Fraction of HTML documents

As an example of the measures representing the fraction of document types used within a site Figure 4.10 provides the distribution of fraction of HTML documents

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Lognormal	$\mu = 1.399$; $\sigma = 0.7673$	0.04081
Brochure	Lognormal	$\mu = 1.349$; $\sigma = 0.4242$	0.01604
Institution	Lognormal	$\mu = 2.107$; $\sigma = 0.4556$	0.01218
Listing	Lognormal	$\mu = 1.678$; $\sigma = 0.5163$	0.01172
Personal	Lognormal	$\mu = 1.382$; $\sigma = 0.3613$	0.01616

Table 4.10: Model distributions and parameters for average level

for each class. Recall that we do not consider embedded objects like pictures in our measurement study. We observe in this figure that more than 80% of the sites from class Personal consist almost entirely, i.e. with more than 95%, of HTML documents. Sites from class Brochure have on average less HTML documents. Here, a large fraction of about 30% of the sites has almost no HTML documents. Another 30% of the sites from Brochure consist of more than 90% of HTML documents. The remaining 40% of these sites are partly composed of HTML and non-HTML documents. The fraction of HTML documents on sites of class Institution is nearly uniformly distributed between 0% and 100%, indicated by the nearly straight line for this class in Figure 4.10. Web sites from the classes Listing and Blog are composed to the least extent of HTML documents. The pages on these sites are mostly generated by PHP scripts. Only 15% or 10% of the sites, respectively, consist almost entirely of HTML documents, and about 70% or 50%, respectively have almost no HTML document on the entire site. In summary, the usage of different document types on the site strongly depends on site class. This is clearly due to the different purposes of the Web sites and the different effort and technical skills of the Web masters. Whereas sites from class Listing, e.g. online shops, mostly generate dynamic pages using Web applications on basis of PHP or other CGI programs, relatively small personal Web sites are in general manually edited by altering the static HTML code. Since sites from class Institution are generally maintained by numerous authors with varying technical knowledge and motivation, the fraction of HTML documents within these sites is nearly uniformly distributed.

For modeling fraction of HTML documents, we distinguish three different

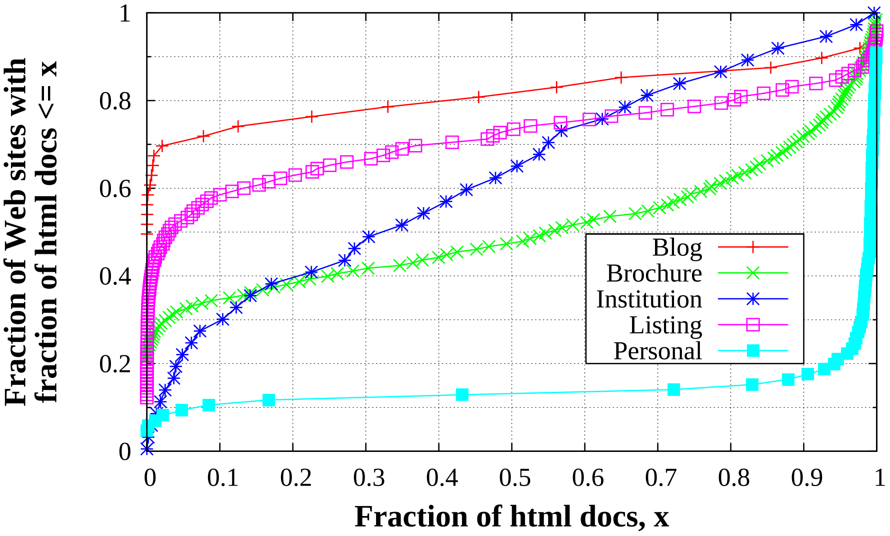


Figure 4.10: Distribution of fraction of HTML documents

regions: The first region is for sites with almost no HTML documents, i.e. with fraction of HTML documents of less than 10%. The second region is for sites with a mixture of HTML and non-HTML documents, i.e. with fraction of HTML documents between 10% and 90%. And finally the third region is for sites which are almost entirely composed of HTML documents, i.e. with fraction of HTML documents of more than 90%. For these regions we specify discrete distributions which reflect the proportions of measured Web sites of each class and region. An exception is made for the class Institution. The distribution of fraction of HTML documents for this class can be approximated by two uniform distributions. One for the first region and another for a combined region comprising the second and the third region. For illustration Figure 4.11 plots the measured and the fitted probability distributions for class Institution. Table 4.11 shows the corresponding parameters of the model distributions for all classes.

4.3.5 Measures Reflecting Composition of URLs

The URLs of Web pages are an easily obtainable but nevertheless meaningful instrument for characterizing the structure of Web sites. Since Web pages are

Class of Web site	**Fitted distribution**	**Matched parameters**
Blog	Discrete	$P\{X \leq 0.1\} = 0.73$
		$P\{0.1 < X \leq 0.9\} = 0.16$
		$P\{X > 0.9\} = 0.11$
Brochure	Discrete	$P\{X \leq 0.1\} = 0.35$
		$P\{0.1 < X \leq 0.9\} = 0.37$
		$P\{X > 0.9\} = 0.28$
Institution	$x \leq 0.1$: Uniform	$P\{X \leq x\} = 3.5 \cdot x$
	$x > 0.1$: Uniform	$P\{X \leq x\} = 0.72 \cdot x + 0.28$
Listing	Discrete	$P\{X \leq 0.1\} = 0.58$
		$P\{0.1 < X \leq 0.9\} = 0.26$
		$P\{X > 0.9\} = 0.16$
Personal	Discrete	$P\{X \leq 0.1\} = 0.11$
		$P\{0.1 < X \leq 0.9\} = 0.06$
		$P\{X > 0.9\} = 0.83$

Table 4.11: Model distributions and parameters for fraction of HTML documents

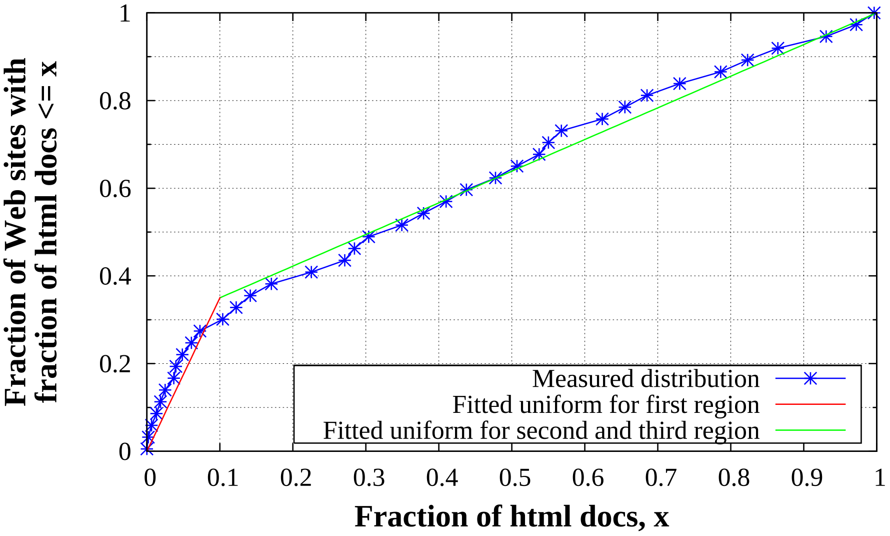


Figure 4.11: Fitting of fraction of HTML documents for class Institution

generally stored in the file system of the Web server the composition of the URLs may indicate the structure of the underlying file system and thus provide insight into the structure of the site. Additionally, the URLs of dynamic Web pages generated by content management systems often contain a significant number of digits, because the URLs do not need to be meaningful for a human who maintains the site. Thus, we analyze a number of structural properties which can directly be extracted from the URLs of a Web site.

Average number of slashes in URL

The first measure reflecting the structure of the underlying file system which stores the Web documents is the average number of slashes found in the URL. A large average number of slashes indicates many subdirectories in the file system and therefore a manually maintained well-structured site. The distribution of this measure for each class is shown in Figure 4.12. We observe that sites from class Institution as expected have on average a larger average number of slashes than sites from other classes. In particular, sites from Personal and Brochure have in general smaller average number of slashes. E.g. only about 10% of these sites

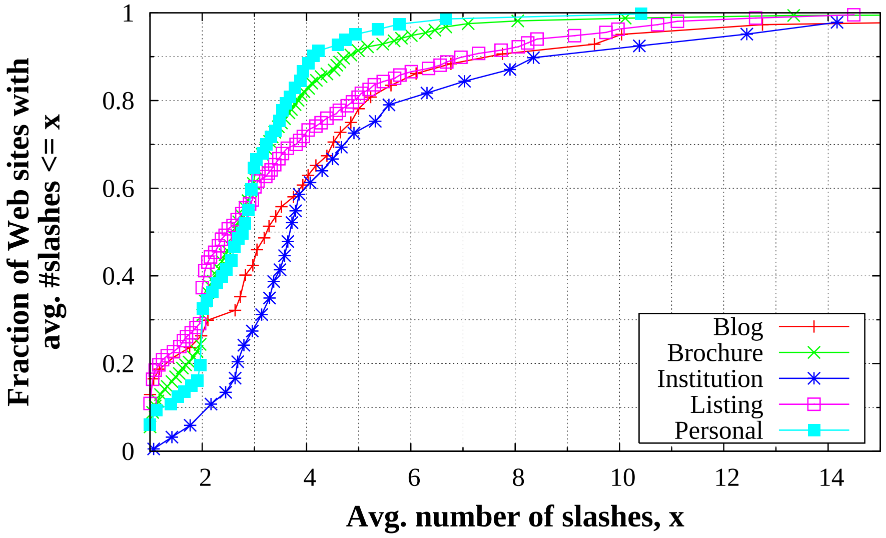


Figure 4.12: Distribution of average number of slashes in URL

have average number of slashes of more than 4. In contrast, 40% of sites from Institution and Blog have this number of slashes. We also find that this measure does not seem to significantly correlate with number of known pages, as could be expected from own experience with managing files in file systems. That is, sites from class Listing have many pages but their URLs contain relatively few slashes. We attribute this behavior to the fact, that sites from class Listing are mostly composed of dynamic documents which are generated from database entries instead of files stored in the file system. This corresponds to the observation made above that only 15% of the sites from class Listing are composed entirely of HTML documents.

In Table 4.12 we provide the model distributions and parameters for average number of slashes. The fitting results show that although the shape of the curves in Figure 4.12 are similar for all classes, the measured distributions are best modeled by three different distribution functions. Whereas the measured distribution for class Blog follows a normal distribution, the measured data follows a Weibull distribution for classes Brochure and Personal. For classes Institution and Listing average number of slashes is best modeled by a lognormal distribution.

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Normal	$\mu = 3.233$; $\sigma = 2.308$	0.02894
Brochure	Weibull	$\alpha = 2.197$; $\lambda = 0.08149$	0.02118
Institution	Lognormal	$\mu = 1.323$; $\sigma = 0.4577$	0.03812
Listing	Lognormal	$\mu = 0.9188$; $\sigma = 0.7872$	0.02151
Personal	Weibull	$\alpha = 2.799$; $\lambda = 0.04457$	0.02850

Table 4.12: Model distributions and parameters for average number of slashes in URL

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Weibull	$\alpha = 2.130$; $\lambda = 0.02123$	0.02177
Brochure	Lognormal	$\mu = 0.8839$; $\sigma = 1.217$	0.02101
Institution	Lognormal	$\mu = 1.749$; $\sigma = 0.9318$	0.02070
Listing	Lognormal	$\mu = 2.394$; $\sigma = 0.9719$	0.01591
Personal	Exponential	$\lambda = 0.2937$	0.01154

Table 4.13: Model distributions and parameters for average number of digits in URL

Average number of digits in URL

As stated above, the number of digits used in the URL provides additional insight into the composition of the Web site. Therefore, we analyze the distributions of average number of digits found in the URL for each class in Figure 4.13. We find that sites from class Listing in general use a larger number of digits in their URLs than sites from the other classes. In contrast, sites from class Personal use on average only few digits. While this behavior can be explained by the different usage of dynamically generated pages, we also observe that Blog sites use relatively few digits in the URL but are composed to a very small extent of HTML documents. Thus, there seems to be no direct correlation between average number of digits in the URL and fraction of HTML documents.

Table 4.13 presents the model distributions and corresponding parameters for

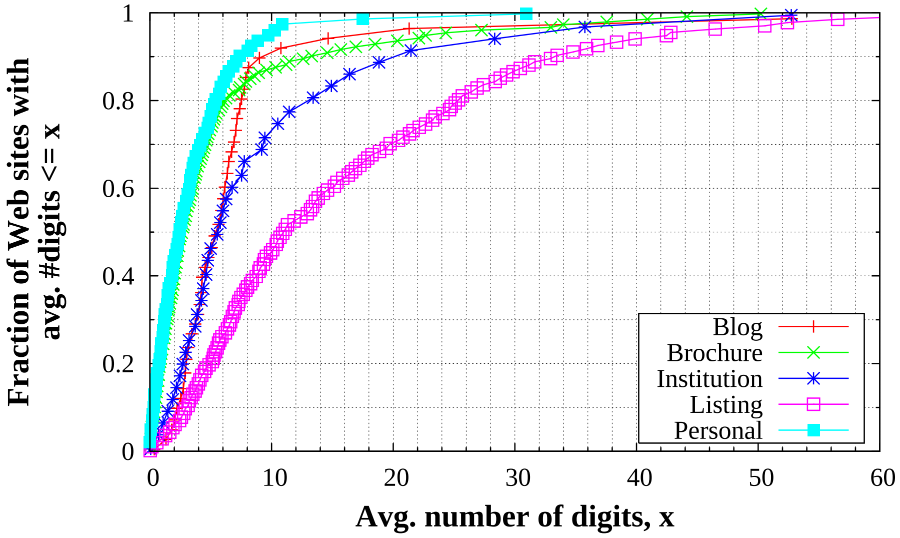


Figure 4.13: Distribution of average number of digits in URL

average number of digits in URL. As in the previous measure, the measured distributions for the individual classes do not follow a common distribution function. Average number of digits follows a Weibull distribution for Blog sites, an Exponential distribution for sites from class Personal and a lognormal distribution for sites from the remaining classes. This different behavior shows that the usage of digits in the URLs is significantly different for the different classes.

Average length of URL path

Finally, as last measure for the characterization of structural properties of Web sites we consider the length of the URL path, i.e. the number of characters in the URL after the host part. Figure 4.14 plots the distributions of average path length for each class. We observe that sites from classes Institution and Listing are similar in this measure and have in general longer URL paths than sites from other classes. Furthermore, sites from classes Blog and Brochure are also similar and have on average longer URL paths than Personal sites. For example, the fraction of sites with average path length of more than 60 characters is about 7% for class Personal, 20% for classes Blog and Brochure, 45% for class Institution

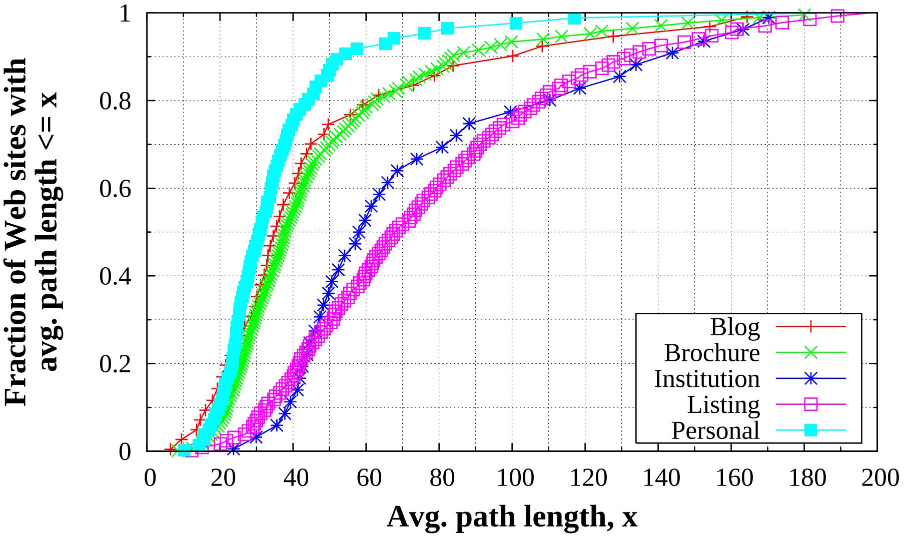


Figure 4.14: Distribution of average length of URL path

and 60% for class Listing. We further observe that short URL paths are very seldom for Institution sites: only about 12% have average path length of less than 40 characters. In contrast, 75% of Personal sites have on average less than 40 characters in their URLs. In fact, a potential correlation between average path length and number of known pages is underlined by the comparison of the distributions for number of known pages (Figure 4.4) and average path length. However, note that sites from class Brochure are similar to Personal sites in number of known pages, whereas these sites are similar to class Blog in average path length. We will discuss the correlation in more detail in the next section.

As for the previous measures, we provide model distributions and parameters for average length of URL path in Table 4.14. The table shows that this measure follows a lognormal distribution for all classes. Furthermore, the relative similarity of the above mentioned groups of classes can be also observed in similar parameters for classes Blog and Brochure, as well as for classes Institution and Listing.

Class of Web site	**Fitted distribution**	**Matched parameters**	**Error term Δ**
Blog	Lognormal	$\mu = 4.099$; $\sigma = 0.3663$	0.03646
Brochure	Lognormal	$\mu = 4.216$; $\sigma = 0.3090$	0.02454
Institution	Lognormal	$\mu = 4.501$; $\sigma = 0.3592$	0.05782
Listing	Lognormal	$\mu = 4.539$; $\sigma = 0.4073$	0.008309
Personal	Lognormal	$\mu = 4.029$; $\sigma = 0.2138$	0.02271

Table 4.14: Model distributions and parameters for average length of URL path

4.4 Correlations between Structural Properties

The previous section demonstrates that the considered structural properties are significantly correlated with the class of a Web site. Since the different classes are categorical data, we showed the correlations by plotting the conditional distributions. For identifying the correlation structure between the individual measures of structural properties we calculate the correlation coefficient for each pair of measures as summarized in Section 2.2. Note that since we already know about the correlation to site classes, the correlation coefficient for each pair of measures is determined for each class separately. By this method we are able to identify differences in the correlation structure between different site classes.

Tables 4.15, 4.16, and 4.17 provide the correlation coefficients for each pair of measures and for each site class. Noting that there is no exact value of the correlation coefficient for distinguishing correlation from independence, we use a value of 0.30 as broad threshold as advised in [HWJ88]. The correlation coefficients exceeding this threshold are marked with a colored background for both positive and negative correlation in all three tables.

From Table 4.15 we observe that the number of known pages has significant influence on most other measures for class Institution. In contrast, number of known pages has less influence on the other structural properties for classes Brochure and Listing. The correlation coefficients indicate that number of known pages correlates to some extent to average number of digits and average path length for all classes. E.g. although the correlation coefficient for number of known pages and average path length calculates to less then the threshold for classes Brochure and Listing, it is only little below the threshold. Thus, these two measures seem not to be completely independent.

The identified correlations between number of known pages and the other measures for classes Institution and Personal indicate that sites from these two classes are to a larger extent maintained manually, because the number of pages of the site has significant impact of how the site is composed. In contrast, Web publishing tools intend to be scalable and thus in general generate site structures which are independent of the number of pages. We further observe correlation between average level and both average number of slashes in URL and average path length for classes Institution and Personal. This supports the before mentioned conjecture that sites from these classes are maintained manually, because the position of the Web pages in the file system corresponds to the composition of the URL.

Table 4.16 shows that average internal outdegree has only little impact on the structure of the Web site. However, for classes Blog and Listing we identify significant correlation of this measure to average document size. That is, the larger the files on a Web sites are, the more links they contain to other pages of the same site. This can be explained by the fact that both Blog and Listing sites present information in an itemized manner. If the elements contain links to other pages, their number increase with increasing number of elements and thus with increasing document size. The same holds for the correlation between average external outdegree and average document size for class Blog. Note that sites from class Listing do not have this correlation because these sites are mostly commercial. Therefore, they avoid to link to external sites. Number of subdomains only correlates to number of known pages for class Institution, clearly because only few sites from other classes use multiple subdomains.

The correlation between average number of slashes and average path length is really obvious for all classes as can be seen in Table 4.17. Since the URL path of a Web page is composed of directory names and the file name of the document containing the page code, a large number of slashes means a large number of directory names and thus a long URL path. However, this correlation is not perfect because long URL paths can also be build by long session IDs or a large number of parameters used for generating dynamic pages. This behavior is expressed in the correlation between average number of digits and average path length which holds for all classes.

Structural properties		Blog	Brochure	Institution	Listing	Personal
number of known pages	average level	0.0132	0.0398	0.3398	-0.0044	0.3067
number of known pages	avg. int. outdegree	0.1225	0.1012	0.3406	0.1112	0.1539
number of known pages	avg. ext. outdegree	0.1098	-0.0073	0.1770	0.0811	0.0100
number of known pages	number of subdomains	0.1135	0.0189	0.3487	0.0000	0.0327
number of known pages	avg. number of slashes	0.3698	0.2579	0.3429	0.1256	0.3973
number of known pages	avg. number of digits	0.1694	0.1168	0.3613	0.2752	0.1449
number of known pages	avg. path length	0.4752	0.2823	0.4669	0.2636	0.5581
number of known pages	fraction of HTML docs	-0.1165	-0.1476	-0.3201	-0.0988	-0.3012
number of known pages	average doc. size	0.1643	-0.0754	-0.2045	0.0113	0.0594
average level	avg. int. outdegree	0.0503	0.0247	-0.0007	-0.0432	-0.1016
average level	avg. ext. outdegree	-0.0101	-0.0328	0.0683	-0.0474	-0.0683
average level	number of subdomains	-0.0090	0.0077	0.1067	-0.0109	0.0404
average level	avg. number of slashes	-0.0812	0.0626	0.3817	0.0530	0.3217
average level	avg. number of digits	-0.0063	0.0637	0.1664	0.0163	0.0972
average level	avg. path length	-0.0640	0.0802	0.4450	0.0875	0.3739
average level	fraction of HTML docs	-0.0666	0.0049	-0.1754	-0.0909	-0.2444
average level	average doc. size	-0.0116	-0.0451	-0.3088	-0.0223	-0.1038

Table 4.15: Correlation coefficient r for correlations to number of known pages and average level

Structural properties		Blog	Brochure	Institution	Listing	Personal
avg. int. outdegree	avg. ext. outdegree	0.2577	0.1865	0.0772	0.1936	0.0549
avg. int. outdegree	number of subdomains	0.2522	0.0151	0.1780	0.0342	0.0484
avg. int. outdegree	avg. number of slashes	0.0333	0.0180	0.0994	0.0925	-0.0690
avg. int. outdegree	avg. number of digits	0.1357	0.0885	0.2342	0.2452	0.1848
avg. int. outdegree	avg. path length	0.08373	0.2063	0.3248	0.2055	0.1258
avg. int. outdegree	fraction of HTML docs	-0.0073	-0.2585	-0.2436	-0.1269	-0.1819
avg. int. outdegree	average doc. size	0.5406	-0.0183	-0.0938	0.5020	0.2190
avg. ext. outdegree	number of subdomains	0.0010	0.0185	0.0789	0.0430	0.2099
avg. ext. outdegree	avg. number of slashes	0.0526	-0.0324	0.2524	-0.0281	-0.0496
avg. ext. outdegree	avg. number of digits	0.0382	0.0545	-0.0621	0.0173	-0.0103
avg. ext. outdegree	avg. path length	0.0036	0.0773	0.1086	-0.0435	-0.0071
avg. ext. outdegree	fraction of HTML docs	-0.0153	-0.0993	-0.0990	0.0349	-0.0698
avg. ext. outdegree	average doc. size	0.5301	-0.0117	-0.0287	0.0794	0.1498
number of subdomains	avg. number of slashes	-0.0424	0.0242	0.0018	0.0412	-0.0394
number of subdomains	avg. number of digits	-0.0437	-0.0006	0.1174	0.0092	0.1252
number of subdomains	avg. path length	-0.0642	0.0093	0.0487	0.0180	0.0148
number of subdomains	fraction of HTML docs	-0.0460	-0.0454	-0.1640	0.0653	-0.1109
number of subdomains	average doc. size	0.1107	-0.0051	-0.0557	-0.0254	-0.0150

Table 4.16: Correlation coefficient r for correlations to avg. internal and external outdegree and number of subdomains

Structural properties		Blog	Brochure	Institution	Listing	Personal
avg. number of slashes	avg. number of digits	0.1613	0.1011	-0.1078	0.2328	0.0645
avg. number of slashes	avg. path length	0.5954	0.5273	0.6528	0.4973	0.6827
avg. number of slashes	fraction of HTML docs	-0.1370	0.0026	-0.1077	-0.1283	-0.1077
avg. number of slashes	average doc. size	0.0759	-0.0690	-0.2029	0.0689	-0.0345
avg. number of digits	avg. path length	0.5724	0.5706	0.4328	0.5983	0.3616
avg. number of digits	fraction of HTML docs	-0.0729	-0.3307	-0.2564	-0.2854	-0.2495
avg. number of digits	average doc. size	0.1791	-0.1034	-0.1657	0.1390	0.0704
avg. path length	fraction of HTML docs	-0.1158	-0.3514	-0.3529	-0.3974	-0.4259
avg. path length	average doc. size	0.1766	-0.1104	-0.3343	0.1127	0.0173
fraction of HTML docs	average doc. size	-0.0387	-0.0822	0.1011	-0.0751	-0.1553

Table 4.17: Correlation coefficient r for correlations to avg. number of slashes and digits in URL, avg. path length, and fraction of HTML documents

We further observe from Table 4.17 that with increasing fraction of HTML documents the path lengths become shorter and to a small extent the number of digits in the URLs decrease. We interpret this behavior such that with increasing fraction of HTML documents, the fraction of dynamic documents using session IDs and other parameters in the URL decrease. Since session IDs etc. are in general relatively large to avoid duplicate assignment, a larger fraction of static HTML documents lead to shorter URL paths.

In summary, we conclude from Tables 4.15, 4.16, and 4.17 that the correlation structure is very different for the different site classes. There are only two pairs of structural properties which are correlated for all classes. All other pairs are either not correlated with sufficient significance or correlated for a subset of the classes only. This indicates, that there seem to be no common rules for composing Web sites, such as "the more pages a site has, the higher the probability of using dynamic documents instead of static HTML documents". This is only true for the classes Institution and Personal.

4.5 Application Example: Classification of Web Sites

The measurement results presented in this chapter can be well applied for the automated coarse classification of Web sites. Such an classifier is currently developed by Lars Littig. However, to keep this thesis self-contained this section outlines how the characterization presented in this chapter can be utilized for classifying Web sites into the five coarse classes Brochure, Listing, Blog, Institution, and Personal.

The naïve Bayesian classifier [DHS01] is known to be a simple but effective technique for classification tasks in several application domains like spam-filtering, text classification, and pattern recognition. As the name "naïve" suggests this method makes the simplifying assumption that the discriminating features, called *discriminators*, are conditionally independent given the class. Although this assumption does not hold in many applications, the naïve Bayesian classifier nevertheless provides excellent classification performance [DP97]. In the application considered here, the disciminators are given by the structural properties of Web sites.

The classifier computes the probability that a Web site belongs to one of the classes Brochure, Listing, Blog, Institution, or Personal, denoted as C_i, with $i = 1, \ldots, 5$ given the set of discriminators (i.e. structural properties) presented in Section 4.3. This probability is denoted as $P(C_i \mid \vec{x})$, where $\vec{x} = \langle x_1, \ldots, x_d \rangle$ is a vector composed of the particular values observed for the discriminators of the Web site to classify, and d is the number of discriminators. The computation is based on several components. First, the likelihood of the discriminators given the considered class, denoted as $P(\vec{x} \mid C_i)$. Second, the prior probability reflecting the fraction of existing Web sites for each class, denoted as $P(C_i)$. Third, a normalizing constant in the denominator, denoted as $P(\vec{x})$, which is invariant across classes. Putting it all together according to Bayes' theorem the probability of a Web site belonging to a specific class given the set of discriminators can be computed by

$$P(C_i \mid \vec{x}) = \frac{P(C_i)P(\vec{x} \mid C_i)}{P(\vec{x})} = \frac{\frac{1}{5}\prod_{j=1}^{d} P(x_j \mid C_i)}{\sum_{i=1}^{5}\left[\frac{1}{5}\prod_{j=1}^{d} P(x_j \mid C_i)\right]} \tag{4.1}$$

Since the fractions of Web sites in the Internet belonging to the particular classes are a priori unknown, we assign for each class the same prior probability, i.e. $P(C_i) = \frac{1}{5}$ for $i = 1, \ldots, 5$. The probabilities $P(x_j \mid C_i)$, $j = 1, \ldots, d$, $i = 1, \ldots, 5$ are given by evaluating the appropriate probability density function $f_{j,i}$ modeling the distribution of discriminator j for class i at location x_j. These probability density functions are provided as model distributions in Section 4.3 for each discriminator and site class. Note that using $f_{j,i}$ in equation 4.1 is not strictly correct because $f_{j,i}$ is a density function rather than a probability. However, the probability for a random variable X lying in some interval is given by $P\{x < X \leq x + h\} = \int_{x}^{x+h} f_{j,i}(x)dx$. Taking into account $\lim_{h \to 0} \frac{P\{x<X\leq x+h\}}{h} = f_{j,i}(x)$ we can approximate $P\{X = x\} \approx f_{j,i}(x) \cdot h$ for some constant h. Since h appears in both the nominator and the denominator in equation 4.1, it cancels out. Thus, equation 4.1 holds for replacing $P(x_j \mid C_i)$ by the probability density functions $f_{j,i}(x_j)$.

By using the model distributions of the structural properties presented in Section 4.3, the naïve Bayes classifier achieves a classification accuracy with a precision of 76% and a recall of 74%. That is, the presented characterization of the structural properties of Web sites allows their automated coarse classification. As mentioned in Section 1.1, this classification beside others enables focused crawling and personalized ranking. Thus, this thesis indirectly contributes to the improvement of Internet search engines.

4.6 Summary

This chapter presented a comprehensive measurement-based study of the German Web. Analyzing structural properties of Web sites, we revealed subtle similarities among functionally related categories of Web sites. In fact, Web sites of the categories Education, Foundation, Medical, and Small and Medium-sized Enterprises (SME) exhibit a similar structure with respect to several measures, e.g. their number of known pages and their average external outdegree. Furthermore, Web sites of the categories City, Forum, Information, and Shop possess significant similarity with respect to their fraction of HTML documents. Subsequently, we aggregated the considered twelve Web site categories into the five classes *Brochure*, *Listing*, *Blog*, *Institution*, and *Personal.*

The characterization of the structural properties of Web sites provides means for an automated classification into these five classes. We observed that number of subdomains constitutes an excellent discriminating measure for Web sites of class Institution. Additionally, Web sites of class Personal can be distinguished from other classes inspecting their fraction of HTML documents. Blog sites exhibit a large external outdegree. In contrast, sites of class Brochure have only few links to external sites. Finally, sites of class Listing have in general many digits in their URLs. Besides highlighting the differences between the considered site classes, we characterized the structural properties by fitting probability distributions to the measured data. We showed that the distributions of structural properties can be well captured by exponential, normal, lognormal, and Weibull probability distributions. Furthermore, a detailed analysis of the correlation structure between the structural properties revealed that site classes do not only differ in the structural properties but also in correlations between these properties.

The characterization presented in this chapter builds the basis for content-independent automatic coarse classification of Web sites. This classification could improve the quality of search results in Internet search engines by enabling focused crawling and personalizes ranking. Furthermore, it supports the manual categorization process of Web directories.

Chapter 5

Concluding Remarks

Due to the dynamic structure and great popularity of P2P file sharing systems and the ever increasing amount of information available in the World Wide Web, a key research challenge for both Internet applications is to develop algorithms for efficiently finding data which is relevant to the user. This thesis supports these research tasks by presenting measurement studies and characterizations of the workload in P2P file sharing systems and of the structural properties of Web sites. The characterization and modeling of the query behavior in P2P file sharing systems is essential for the development and evaluation of novel search protocols. Moreover, exploiting the structural properties of Web sites enables the development of new Internet search engine technology which significantly improves the quality of search results.

Characterization of the Query Behavior in Peer-to-Peer File Sharing Systems

As a building block for the performance evaluation of P2P file sharing systems, this thesis provided a detailed synthetic workload model for the query behavior of peers in a P2P file sharing system. In order to obtain representative data of the user behavior in current P2P file sharing systems, we conducted a comprehensive measurement study of the messages exchanged in the Gnutella overlay network which is one of the most popular P2P file sharing systems today. Our analysis approach has three unique key features. (1) We only considered queries which were issued by direct neighbors in the Gnutella overlay topology. By this, each query is uniquely related to its issuer. (2) We filtered out queries which certain client implementation automatically issued in order to improve responsiveness of the system. Those automated queries are e.g. re-queries with identical query string which are issued to find more sources for continuing a running download. To identify those automated queries, this thesis presented five filter rules implementing appropriate heuristics. Applying these filter rules, we separated pure user behavior from system-specific features. (3) By relating each query to the corresponding issuer, we analyzed the query workload of individual peers opposed to aggregate measures presented in previous work. This enabled for each query the identification of the geographical region, the time-of-day, and other user session related measures.

For deriving a synthetic workload model, this thesis characterized all workload measures required to model the user-induced query behavior. These measures include the fraction of connected sessions that are passive, i.e. which do not include user-induced queries, the duration of such sessions, and for each active session, the number of queries issued, the interarrival times between issuing two subsequent queries, the time between session start and issuing the first query, the time between issuing the last query and end of session, and the popularity of the issued queries. For all measures, the proposed workload model captures the key correlations to the geographical region of the peer, the time-of-day, and to other workload measures, such that realistic synthetic workload can be generated. The characterization revealed the following key observations:

1. Automated re-queries generated by the client software have a significant impact on most workload measures. This observation shows on the one hand that these queries have to be filtered out to characterize user behavior. On the other hand it indicates potential for improving the Gnutella search algorithm to make automated re-queries obsolete.

2. The 100 most popular queries change significantly from one day to the next. That is, in a measurement study over multiple days, the relative popularity of individual queries may be low, while the relative popularity for the same queries may be high considering a single day only. Thus, realistic synthetic workload models for generating detailed query behavior must not be based on the aggregate query popularity derived from a multiple-day measurement.

3. 97% of the queries issued from peers in North America are not issued by peers in Europe, and vice versa. This observation may have significant impact on the design of future P2P systems, in that it shows that the interests of peers in different geographical regions differ significantly (at least at a specific point in time). That is, the performance of P2P file sharing systems may be significantly improved by taking the geographical region of the peers into consideration. For example, queries could be directed only to peers in the same geographical region, because there the interest in the same file, and therewith the probability of getting results, is much stronger.

4. The number of queries per session and passive session duration are also sensitive to geographical region, with peers in Europe issuing more queries per session and having longer passive session durations, on average. Again this observation underlines the importance of distinguishing between peers from different geographical regions when generating realistic workload.

5. The time between the last query and the end of the session has a much heavier tail than the time between queries. That is, peers tend to remain passively connected to the Gnutella network after issuing their last query much longer than they are passive between two subsequent queries. This indicates, that, in general, peers connect to the P2P system, issue a number of queries for finding the files of interest and then stay connected passively while downloading the found files.

As stated above, the analysis presented in this thesis showed that there are substantial differences in the query behavior of peers in different geographical regions. Similarly, Le Fessant et al. identified geographical as well as interest-based locality in the files shared by the peers [FHKM04]. Sripanidkulchai et al. additionally found interest-based locality in download requests of KaZaA and Gnutella peers [SMZ03]. While these studies provide initial indication for interest-based or geographical clustering of peers, the degree and implications of this phenomenon is subject to future research directions.

A second future research field arises from the increasing prevalence of mobile devices such as notebooks, personal digital assistants, or smart phones. Equipped with wireless network technology these devices may form self-configuring Mobile Ad Hoc Networks (MANET) which enable multi-hop communication without central infrastructure, or connect to the Internet via base stations. Although these mobile networks gain increasing popularity the characteristics and performance of P2P systems run over those networks are entirely unknown, thus constituting an additional research challenge.

Characterization of Web Sites by their Structural Properties

As second main contribution, this thesis identified substantial differences in the structural properties of Web sites which allow for their automated classification. The automated classification of Web sites enables novel functionality of Internet search engines like focused crawling and personalized ranking. Thus, this thesis indirectly contributes to the improvement of Internet search engine technology.

The characterization of Web sites presented in this thesis is based on a comprehensive measurement study of the German part of the Web. The study comprises more than 2,300 Web sites consisting of more than 11 million Web pages and reflect the twelve major categories of Web sites. These categories include, amongst others, academic sites, Web logs, information portals, online shops, the Web presence of small and medium -sized enterprises (SME), as well as private homepages. Thus, the analysis is based on a representative data set for the German Web.

As first key observation, our analysis revealed that Web sites of functional related categories exhibit similar structure with respect to several measures. These measures reflect the size, the organization, the composition of URLs and the link structure of the sites. For example, Web sites belonging to the categories Education, Medical, Foundation, and SME have the same function in that they present information about a specific institution. Similarly, sites from the categories City, Forum, Information, and Shop list a large number of items, i.e. products in the case of online shops or news in the case of information portals. These sites are also similar in their number of pages and their average external outdegree, i.e. in their structural properties. Furthermore, Web sites of the categories City, Forum, Information, and Shop possess significant similarities with respect to their fraction of HTML documents. These similarities allowed the aggregation of the twelve Web site categories into the five classes *Brochure*, *Listing*, *Blog*, *Institution*, and *Personal* based on both, structural properties and function.

Based on these five classes of Web sites, this thesis presented a detailed characterization of the structural properties of Web sites belonging to the different classes. Key observations of this characterization include:

1. The number of known pages is an excellent measure for distinguishing between sites of the classes Brochure and Personal (which are very similar in this property) from sites of the classes Listing and Institution (which are also similar in this measure). That is, sites of the former group have on average fewer pages than sites of the latter group. The number of pages of Web log sites lies on average between the two groups.

2. The average document size distinguishes sites belonging to class Personal from sites of class Blog and sites of the other classes. On average, sites from class Personal have the smallest document sizes, i.e. more than 85% of these sites have an average document size of less than 20 kBytes. In contrast, 37% of the Web logs have an average document size of more than 20 kBytes. Sites from the other classes exhibit even larger documents on average. Since the document size is measured without considering embedded objects like pictures, this result shows that private homepages and Web logs provide less information per page than the sites of the other classes. Combined with the results obtained for the number of pages, our analysis verifies that private homepages and Web logs are typically small compared to the other sites.

3. The pages of sites belonging to class Listing have many internal outlinks, whereas pages of sites from class Personal link to only few other pages on the same site. While it could be expected that this behavior is caused by the different number of pages (the more pages you have, the more pages you can link to), the correlation analysis of the number of known pages and average internal outdegree shows no significant correlation for all classes but Institution.

4. Blog sites exhibit a large external outdegree while sites of class Brochure have only few links to external sites. This is clearly caused by the different purpose of the sites. While Web logs are destined for providing strong link structure to other Web logs, sites of class Brochure try to prevent a surfer from leaving the site by not linking to potential competitors.

5. Web sites of class Personal can be distinguished from other classes inspecting their fraction of HTML documents. That is, private homepages most often completely consist of static HTML documents, while sites from the

other classes are typically at least partly composed of dynamic HTML documents generated by PHP, ASP, JSP, or some other server-side scripts.

6. Listing sites use on average more digits in the URLs. Since the items listed on sites of class Listing such as news or products are more dynamic than the information presented e.g. on private homepages or institutional sites, the pages are typically generated using certain item IDs. Furthermore, many of these sites, in particular online shops, make extensive use of session IDs. Thus, the IDs lead to large number of digits in the URLs.

7. The average path length significantly correlates to the number of slashes and the number of digits for all classes of Web sites. This shows that long URL paths are caused by URLs with many slashes, i.e. a deep tree structure of the Web server's file system, and by URLs with many digits, i.e. long session or item IDs. Other significant correlations between structural properties only hold for subsets of site classes. Thus, sites belonging to different classes not only differ in the values of structural properties but also in the correlations between them.

The results gained in this thesis open up a number of future research directions. Exploiting the differences in the structural properties for developing an automated classifier is a straightforward extension of this thesis. Initial experiments with a naïve Bayesian approach conducted by Lars Littig already yielded classification accuracy with a precision of 76% and a recall of 74%. Nevertheless, the development of more advanced classifiers based on the structural properties of Web sites may likely achieve even higher classification accuracy. Furthermore, the comparison of the structural properties of Web sites in other geographical regions with those from Germany presented in this thesis will provide additional understanding of geographical influences. Finally, since the Internet is changing with incredible velocity (suppose e.g. the rapid advent of Web logs) an important question is, how the structural properties change over time. The examination of the evolution of Web site structure would foster the insight into the forces that drive the Web.

Bibliography

[ACD$^+$03] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The Connectivity Sonar: Detecting Site Functionality by Structural Patterns. In *Proc. HyperText 2003*, Nottingham, United Kingdom, August 2003.

[ACGM$^+$01] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. *ACM Trans. on Internet Technology*, 1:2–43, 2001.

[AH00] E. Adar and B. Huberman. Free Riding on Gnutella. Technical report, Xerox PARC, 2000.

[BC98] P. Barford and M. Crovella. Generating Representative Web Workloads for Network and Server Performance Evaluation. In *Proc. ACM Sigmetrics*, pages 151–160, July 1998.

[BC99] P. Barford and M. Crovella. Measuring Web Performance in the Wide Area. *Performance Evaluation Review*, Special Issue on Network Traffic Measurement and Workload Characterization, August 1999.

[Bit04] BitTorrent, Inc. BitTorrent Homepage. `http://www.bittorrent.com`, October 2004.

[BKM$^+$00] A. Broder, R. Kumar, F. Maghoul, P. Rhaghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph Structure in the Web. In *Proc. Int. WWW Conf. (WWW 2000)*, Amsterdam, The Netherlands, May 2000.

[BRVX04] L. Bent, M. Rabinovich, G. Voelker, and Z. Xiao. Characterization of a Large Web Site Population with Implications for Content Delivery. In *Proc. Int. WWW Conf. (WWW 2004)*, New York, NY, May 2004.

[BSV03] R. Bhagwan, S. Savage, and G. Voelker. Understanding Availability. In *Proc. Int. Workshop on P2P Systems (IPTPS 2003)*, Berkeley, CA, February 2003.

[BW88] D. Bates and D. Watts. *Nonlinear Regression Analysis and Its Applications.* Wiley, 1988.

[CB97] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, December 1997.

[CG04] L. Cherkasova and M. Gupta. Analysis of Enterprise Media Server Workloads: Access Patterns, Locality, Content Evolution, and Rates of Change. *IEEE/ACM Transactions on Networking*, 12(5):781–794, October 2004.

[CGM00] J. Cho and H. Garcia-Molina. The Evolution of the Web and its Implications for an Incremental Crawler. In *Proc. Int. Conf. on Very Large Data Bases (VLDB 2000)*, Cairo, Egypt, September 2000.

[CLL02] J. Chu, K. Labonte, and B. Levine. Availability and Locality Measurements of Peer-to-Peer File Systems. In *Proc. SPIE ITCom: Scalability and Traffic Control in IP Networks*, Boston, MA, July 2002.

[Coo06] Cooperative Association for Internet Data Analysis - CAIDA. Internet Measurement Infrastructure. `http://www.caida.org/analysis/performance/measinfra/`, January 2006.

[CRB+03] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker. Making Gnutella-like P2P Systems Scalable. In *Proc. ACM Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM 2003)*, Karlsruhe, Germany, August 2003.

[CV01] S. Chiang and M. Vernon. Characteristics of a Large Shared Memory Production Workload. In *Proc. Workshop on Job Scheduling Strategies for Parallel Processing*, Cambridge, MA, June 2001.

[DHS01] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* John Wiley and Sons, 2nd edition, 2001.

[DKM+02] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-Similarity in the Web. *ACM Trans. on Internet Technology*, 2:205–223, 2002.

[DP97] Pedro Domingos and Michael J. Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning*, 29(2-3):103–130, 1997.

[Edw76] A. Edwards. *An Introduction to Linear Regression and Correlation*, chapter The Correlation Coefficient, pages 33–46. W.H. Freeman, San Francisco, CA, 1976.

[EKL05] S. ElRakabawy, A. Klemm, and C. Lindemann. TCP with Adaptive Pacing for Multihop Wireless Networks. In *Proc. ACM Int. Symp. on Mobile Ad Hoc Networking and Computing (Mobihoc 2005)*, Urbana-Champain, IL, May 2005.

[EKS02] M. Ester, H. Kriegel, and M. Schubert. Web Site Mining: A New Way to Spot Competitors, Customers and Suppliers in the World Wide Web. In *Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD 2002)*, Edmonton, Canada, July 2002.

[FHKM04] F. Le Fessant, S. Handurukande, A. Kermarrec, and L. Massoulie. Clustering in Peer-to-Peer File Sharing Workloads. In *Proc. Int. Workshop on Peer-to-Peer Systems (IPTPS 2004)*, San Diego, CA, February 2004.

[FML+03] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and S. Diot. Packet-level Traffic Measurements from the Sprint IP Backbone. *IEEE Network Magazine*, 17(6):6–16, November-December 2003.

[FMNW03] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A Large-scale Study of the Evolution of Web Pages. In *Proc. Int. WWW Conf. (WWW 2003)*, Budapest, Hungary, May 2003.

[GDS$^+$03] K. Gummadi, R. Dunn, S. Saroiu, S. Gribble, H. Levy, and J. Zahorjan. Measurement, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload. In *Proc. ACM Symp. on Operating Systems Principles (SOSP 2003)*, Bolton Landing, NY, October 2003.

[GFJ$^+$03] Z. Ge, D. Figueiredo, S. Jaiswal, J. Kurose, and D. Towsley. Modeling Peer-Peer File Sharing Systems. In *Proc. Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM 2003)*, San Francisco, CA, March-April 2003.

[GGM05] Z. Gyöngyi and H. Garcia-Molina. Spam: It's Not Just for Inboxes Anymore. *IEEE Computer*, 38(10):28–34, 2005.

[Gnu04] Gnutella Developer Forum. Gnutella - A Protocol for a Revolution. `http://rfc-gnutella.sourceforge.net`, October 2004.

[Goo05] Google, Inc. Google Homepage. `http://www.google.com`, October 2005.

[GPT05] D. Gibson, K. Punera, and A. Tomkins. The Volume and Evolution of Web Page Templates. In *Proc. Int. WWW Conf. (WWW 2005)*, Chiba, Japan, May 2005.

[GTL$^+$02] E. Glover, K. Tsioutsiouliklis, S. Lawrence, D. Pennock, and G. Flake. Using web structure for classifying and describing web pages. In *Proc. Int. WWW Conf. (WWW 2002)*, Honolulu, Hawaii, May 2002.

[Hop05] Hoppenstedt GmbH. Hoppenstedt Homepage. `http://www.hoppenstedt.de`, October 2005.

[HWJ88] D. Hinkle, W. Wiersma, and S. Jurs. *Applied Statistics for the Behavioral Sciences.* Houghton Mifflin College Div, 2 edition, 1988.

[ISZ99] A. Iyengar, M. Squillante, and L. Zhang. Analysis and Characterization of Large-Scale Web Server Access Patterns and Performance. *World Wide Web*, 2(1-2):85–100, 1999.

[JKB94] N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, 2 edition, 1994.

[KK62] J. Kenney and E. Keeping. *Mathematics of Statistics*, chapter Root Mean Square, pages 59–60. Van Nostrand, Princeton, 3 edition, 1962.

[KLL01] A. Klemm, C. Lindemann, and M. Lohmann. Traffic Modeling and Characterization for UMTS Networks. In *Proc. Globecom 2001, Internet Performance Symposium*, San Antonio, TX, November 2001.

[KLL02] A. Klemm, C. Lindemann, and M. Lohmann. Traffic Modeling of IP Networks Using the Batch Markovian Arrival Process. In *Proc. Int. Conf. on Modeling Tools and Techniques for Computer and Communication System Performance Evaluation (Tools 2002)*, volume 2324. Lecture Notes in Computer Science, Springer-Verlag, April 2002.

[KLL03] A. Klemm, C. Lindemann, and M. Lohmann. Modeling IP Traffic Using the Batch Markovian Arrival Process. *Performance Evaluation*, 54:149–173, 2003.

[KLVW04] A. Klemm, C. Lindemann, M. Vernon, and O. Waldhorst. Characterizing the Query Behavior in Peer-to-Peer File Sharing Systems. In *Proc. ACM Internet Measurement Conference (IMC 2004)*, Taormina, Italy, October 2004.

[KLW03] A. Klemm, C. Lindemann, and O. Waldhorst. A Special-Purpose Peer-to-Peer File Sharing System for Mobile Ad Hoc Networks. In *Proc. IEEE Semiannual Vehicular Technology Conference (VTC2003-Fall)*, Orlando, FL, October 2003.

[KLW04a] A. Klemm, C. Lindemann, and O. Waldhorst. Peer-to-Peer Computing in Mobile Ad Hoc Networks. In M. Calzarossa and E. Gelenbe, editors, *Performance Tools and Applications to Networked Systems*, volume 2965 of *Lecture Notes in Computer Science*. Springer-Verlag, 2004.

[KLW04b] A. Klemm, C. Lindemann, and O. Waldhorst. Relating Query Popularity and File Replication in the Gnutella Peer-to-Peer Network. In *Proc. GI/ITG Conference on Measuring, Modeling and Evaluation of Computer and Communication Systems (MMB 2004)*, Dresden, Germany, September 2004.

[KS04] H. Kriegel and M. Schubert. Classification of Websites as Sets of Feature Vectors. In *Proc. Int. Conf. on Databases and Applications (DBA 2004)*, Innsbruck, Austria, February 2004.

[KWX01] B. Krishnamurthy, J. Wang, and Y. Xie. Early Measurements of a Cluster-Based Architecture for P2P Systems. In *Proc. ACM Sigcomm Internet Measurement Workshop (IMW 2001)*, San Francisco, CA, November 2001.

[LCC+02] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. Search and Replication in Unstructured Peer-to-Peer Networks. In *Proc. ACM Int. Conf. on Supercomputing*, New York, NY, June 2002.

[Loh04] M. Lohmann. *Online QoS/Revenue Management for Third Generation Mobile Communication Networks.* PhD thesis, University of Dortmund, Department of Computer Science, Dortmund, Germany, 2004.

[LTK+00] C. Lindemann, A. Thümmler, A. Klemm, M. Lohmann, and O. Waldhorst. Quantitative System Evaluation with DSPNexpress 2000. In *Proc. Int. Workshop on Software and Performance (WOSP 2000)*, Ottawa, Canada, September 2000.

[LTK+02] C. Lindemann, A. Thümmler, A. Klemm, M. Lohmann, and O. Waldhorst. Performance Analysis of Time-enhanced UML Diagrams Based on Stochastic Processes. In *Proc. Int. Workshop on Software and Performance (WOSP 2002)*, Rome, Italy, July 2002.

[Max04] MaxMind, LLC. Geotargeting IP Address. `http://www.maxmind.com`, October 2004.

[Met05] MetaMachine. eDonkey2000 Homepage. http://www.edonkey2000.com, October 2005.

[MEW00] A. Mahanti, D. Eager, and C. Williamson. Temporal Locality and its Impact on Web Proxy Cache Performance. *Performance Evaluation*, 42:187–203, 2000.

[Mic05] Microsoft Corporation. MSN Search. http://search.msn.com, December 2005.

[Mil92] D. Mills. Network Time Protocol (Version 3) Specification, Implementation and Analysis. RFC 1305, March 1992. available at http://www.ietf.org/rfc/rfc1305.txt.

[Min06] Miniwatts Marketing Group. Internet Growth Statistics. http://www.internetworldstats.com/stats.htm, January 2006.

[MS97] S. Manley and M. Seltzer. Web Facts and Fantasy. In *Proc. USENIX Symposium on Internet Technologies and Systems (USITS 1997*, Monterey, CA, December 1997.

[Mut04] Mutella. Mutella Homepage. http://mutella.sourceforge.net, October 2004.

[NCO04] A. Ntoulas, J. Cho, and C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In *Proc. Int. WWW Conf. (WWW 2004)*, New York, NY, May 2004.

[NCR$^+$03] E. Ng, Y. Chu, S. Rao, K. Sripanidkulchai, and H. Zhang. Measurement-Based Optimization Techniques for Bandwidth-Demanding Peer-to-Peer Systems. In *Proc. Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM 2003)*, San Francisco, CA, March-April 2003.

[Net05] Netscape Communications Corporation. DMOZ: open directory project. http://www.dmoz.org, October 2005.

[Net06] Netcraft LTD. Netcraft: Web Server Survey. http://news.netcraft.com/archives/web_server_survey.html, January 2006.

[Pax04] V. Paxson. Strategies for Sound Internet Measurements. In *Proc. ACM Internet Measurement Conference (IMC 2004)*, Taormina, Italy, October 2004.

[Pit99] J. Pitkow. Summary of WWW Characterizations. *World Wide Web*, 2(1-2):3–13, 1999.

[RFH$^+$01] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A Scalable Content-Addressable Network. In *Proc. ACM Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM 2001)*, San Diego, CA, August 2001.

[RIF02] M. Ripeanu, A. Iamnitchi, and I. Foster. Mapping the Gnutella Network. *IEEE Internet Computing*, pages 50–57, January-February 2002.

[SGD$^+$02] S. Saroiu, K. P. Grummadi, R. J. Dunn, S. D. Gribble, and H. M. Levy. An Analysis of Internet Content Delivery Systems. In *Proc. 5th Symposium on Operating Systems Design and Implementation (OSDI 2002)*, Boston, MA, December 2002.

[SGG02] S. Saroiu, K. Gummadi, and S. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *Proc. Conf. on Multimedia Computing and Networking (MMCN 2002)*, San Jose, CA, January 2002.

[Sha04] Sharman Networks Ltd. KaZaA Media Desktop. `http://www.kazaa.com`, October 2004.

[SMK$^+$01] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In *Proc. ACM Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM 2001*, San Diego, CA, August 2001.

[SMZ03] K. Sripanidkulchai, B. Maggs, and H. Zhang. Efficient Content Location Using Interest-Based Locality in Peer-to-Peer Systems. In

Proc. Joint Conf. of the IEEE Computer and Communications Societies (INFOCOM 2003), San Francisco, CA, March-April 2003.

[Sri01] K. Sripanidkulchai. The Popularity of Gnutella Queries and its Implications on Scalability. Featured on O'Reilly's `http://www.openp2p.com/topics/p2p/gnutella/` Web site, February 2001.

[SRS05] D. Stutzbach, R. Rejaie, and S. Sen. Characterizing Unstructured Overlay Topologies in Modern P2P File-Sharing Systems. In *Proc. ACM Internet Measurement Conference (IMC 2005)*, Berkely, CA, October 2005.

[SW02] S. Sen and J. Wang. Analyzing P2P Traffic Across Large Networks. In *Proc. Internet Measurement Workshop (IMW 2002)*, Marseilles, France, November 2002.

[THG$^+$03] Y. Tian, T. Huang, W. Gao, J. Cheng, and P. Kang. Two-Phase Web Site Classification Based on Hidden Markov Tree Models. In *IEEE/WIC Int. Conf. on Web Intelligence (WI'03)*, Hallifax, Canada, October 2003.

[Thü03] Axel Thümmler. *Stochastic Modeling and Analysis of 3G Mobile Communication Systems.* PhD thesis, University of Dortmund, Department of Computer Science, Dortmund, Germany, 2003.

[Tru05] Trusted Shops GmbH. Trusted Shops. `http://www.trustedshops.de`, October 2005.

[Wal05] O. Waldhorst. *Design and Quantitative Analysis of Protocols for Epidemic Information Dissemination in Mobile Ad Hoc Networks.* PhD thesis, University of Dortmund, Department of Computer Science, Dortmund, Germany, 2005.

[WK05] T. Williams and C. Kelley. Gnuplot Homepage. `http://www.gnuplot.info`, October 2005.

[Yah05] Yahoo!, Inc. Yahoo Homepage. `http://www.yahoo.com`, October 2005.

[YGM01] B. Yang and H. Garcia-Molina. Comparing Hybrid Peer-to-Peer Systems. In *Proc. Int. Conf. on Very Large Data Bases (VLDB 2001)*, Rome, Italy, September 2001.

[ZSR06] S. Zhao, D. Stutzbach, and R. Rejaie. Characterizing Files in the Modern Gnutella Network: A Measurement Study. In *Proc. Multimedia Computing and Networking (MMCN 2006)*, San Jose, CA, January 2006.