

A Characterization of Broadband User Behavior and Their E-Business Activities

Humberto T. Marques Neto
Jussara M. Almeida

Leonardo C. D. Rocha
Wagner Meira Jr.

Pedro H. C. Guerra
Virgilio A. F. Almeida

{hmarques, lcrocha, pcalais, jussara, meira, virgilio}@dcc.ufmg.br¹

Abstract

This paper presents a characterization of broadband user behavior from an Internet Service Provider standpoint. Users are broken into two major categories: residential and Small-Office/Home-Office (SOHO). For each user category, the characterization is performed along four criteria: (i) session arrival process, (ii) session duration, (iii) number of bytes transferred within a session and (iv) user request patterns.

Our results show that both residential and SOHO session inter-arrival times are exponentially distributed. Whereas residential session arrival rates remain relatively high during the day, SOHO session arrival rates vary much more significantly during the day. On the other hand, a typical SOHO user session is longer and transfers a larger volume of data. Furthermore, our analysis uncovers two main groups of session request patterns within each user category. The first group consists of user sessions that use traditional Internet services, such as e-mail, instant messenger and, mostly, www services. On the other hand, sessions from the second group, a smaller group, use typically peer-to-peer file sharing applications, remain active for longer periods and transfer a large amount of data. Looking further into the e-business services most commonly accessed, we found that subscription-based and advertising services account for the vast majority of user HTTP requests in both residential and SOHO workloads. Understanding these user behavior patterns is important to the development of more efficient applications for broadband users.

1 Introduction

Understanding the nature and characteristics of broadband user behavior is a crucial step to improve the quality of service offered to users in the next generation broadband environments. Broadband user behavior characterization can lead to a better understanding of the interaction between users and service providers. It can also help the design of systems with better QoS metrics, such as performance, availability, security and cost.

Broadband penetration keeps growing fast for users and

¹Department of Computer Science, Federal University of Minas Gerais, Brazil - 31270-010

households. However, studies of broadband user behavior are scarce in the literature, mainly because of the difficulty in obtaining actual logs from Internet Service Providers (ISPs). Most of the service providers on the Internet consider logs as very sensitive data. Existing studies, such as the one done by Pew Internet & American Life [2], concentrate on qualitative analysis based on surveys. The Pew report shows how on-line Americans' behavior changes with high speed connections at home. The study also shows that broadband services allow users to distinguish themselves from dial-up counterparts in the following ways: (i) broadband users engage in multiple Internet activities on a daily basis, (ii) high speed users become creators and managers of different types of on-line content and (iii) broadband users perform a large variety of queries for information. In spite of the Pew report, quantitative studies of broadband user behavior are still lacking.

This paper intends to fill this gap. To understand the broadband user behavior, we present a characterization from a broadband ISP (a TV cable company that provides broadband services to its users), which classifies their users into two major categories: residential and Small-Office/Home-Office (SOHO). For each category, we identify user sessions, which are defined as the period that an user is connected to the broadband network. Basically, the behavior of users is defined as a function of the way users arrive at the ISP, how long they remain on-line, the number of bytes they transfer and what they do while connected, i.e., the request pattern within a session. Thus, the characterization process is performed along four criteria: (i) session arrival process, (ii) session duration, (iii) number of bytes transferred within a session and (iv) user request pattern. The broadband user behavior characterization is based on logs collected on an authentication server and by Netflow [1] running in a border router. The data collecting architecture implemented in the ISP allows us to identify the services used by each user category.

In order to analyze the service request patterns, we use a state transition graph called Customer Behavior Model Graph (CBMG) [14], which describes the behavior of groups of customers who exhibit similar navigational patterns. We then applied clustering algorithms to user session data (both residential and SOHO) to determine groups of users that exhibit similar behavior graphs. Finally, we look further into the HTTP-based web services most frequently accessed by

the residential and SOHO broadband users in our workload and characterize the popularity of different categories of e-business services.

The main findings of the characterization study are:

- Both residential and SOHO session inter-arrival times are exponentially distributed during periods of stable arrival rates. Residential session arrival rates remain relatively high during the day, whereas SOHO session arrival rates vary significantly through the day.
- Residential user session durations can be accurately approximated by a Lognormal distribution. On the other hand, the duration of SOHO user sessions is better modeled with a combination of a Lognormal distribution for the body and a Pareto for the tail.
- For both residential and SOHO sessions, the numbers of incoming and outgoing bytes can be modeled with Lognormal distributions. In addition to that, the typical ratio of the average number of incoming bytes to the average number of outgoing bytes per session falls in the 3 to 5 range.
- The use of a state transition graph (CBMG) uncovered six classes of significantly different patterns in the user behavior of both categories. For example, one class is dominated by HTTP requests. A second class has a lot of HTTP requests but include some requests to other services, such as P2P, Instant Messengers etc. Another class is dominated by P2P requests, which lasts much longer than HTTP sessions.
- The vast majority of the e-business services requested by both residential and SOHO users in our workloads (79 out of 164 classified web sites) provide content and services either on a subscription basis or mixed with advertisements.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the data collection process and the characterization methodology. Section 4 presents the most relevant results from our characterization. Concluding remarks are offered in Section 5.

2 Related Work

A number of workload and user behavior characterizations are available in the literature. Traditional web workloads, consisting mainly of HTTP requests to documents and image files, are analyzed in several previous studies, focused on either server-side [3, 5] or client-side workloads [6, 8].

More recent studies characterize the workloads of other types of applications like on-demand and live distribution of streaming media [7, 22] and peer-to-peer (P2P) services

[10, 13, 18, 19], which are both becoming increasingly popular, possibly due to the availability of broadband “last mile” connections [12]. Previous streaming media workload characterizations propose hierarchical models to capture the most relevant aspects of user behavior for the specific type of workload studied (live [22] and on-demand [7]) and produce extensive characterizations of each component of the proposed models. Previous peer-to-peer workload studies analyze several aspects of the traffic generated by these applications such as object popularity, object size, bandwidth utilization and session durations [10, 11, 13, 18, 19].

These previous workload analysis focus on a specific type of application. In contrast, our work looks into a client-side broadband workload including requests to a multitude of different applications. In that sense, the extensive characterization of a broadband ISP web proxy [4] is possibly the previous work that is most closely related to ours. However, that work focuses on the traffic generated by the analyzed broadband community, characterizing file types, sizes, popularity and frequency of requests to different services (i.e., HTTP, FTP etc). In contrast, our focus is not only on the traffic generated by broadband users but, especially, on the patterns of user requests to different services that most accurately represent the typical behavior within a broadband session. In other words, we characterize not only traditionally analyzed aspects such as user session arrival process, duration and traffic volume but also the most commonly observed patterns of user requests to different services within the same session. Furthermore, we also contrast our findings for two different categories of broadband users: residential and SOHO users.

3 Characterization Methodology

This section presents our characterization methodology and describes how it is applied to the ISP environment. The goal is to analyze the user activity while connected to the Internet, quantifying and qualifying the workload they generate.

Our characterization is based on four criteria: session arrival process, session duration, traffic volume, and user request pattern. The session arrival process and session duration provide temporal information about the workload generated by the users, since we may estimate how frequently and for how long a user is connected. The traffic volume provides leverage on how the users are using their connection regarding a critical resource for any ISP: bandwidth. Finally, the user request pattern qualifies the nature of the services being requested by users and how they are distributed across the connection time.

We employ three sources of data in the proposed characterization: user authentication log, user database, and traffic log. The user authentication log is compatible with the RADIUS protocol [16, 17]. It has an entry for each user session containing the following information: start date and time, duration, number of bytes transferred, and the dynamic IP as-

signed to the user. The user database is a table that informs the user category, namely residential or Small-Office/Home-Office (SOHO), in the ISP analyzed. The third log is collected using Netflow [1]. The traffic is divided into flows and each flow is characterized by a timestamp that indicates when the flow was recorded, source and destination IPs, protocols and ports, and the volume of bytes transferred. The traffic log, from which the results presented in the next section are generated, was collected at one of the three backbone routers of the ISP and corresponds to about 30% of the overall traffic. Since the user population is equally spread across the three routers, we believe that the data gathering at a single router does not affect the statistical meaning of the results. In other words, the user population analyzed is representative of the ISP user community.

Before characterizing the workload along each criterion, we divide the data into two sets according to the user categories, defining, thus, two separate broadband workloads. Our *residential workload* consists of all sessions initiated by users categorized as such by the service provider. Similarly, the *SOHO workload* consists of all SOHO user sessions. We then characterize the four user behavior criteria for each workload.

The session arrival, session duration and traffic volume criteria are characterized by using the authentication log. For the sake of characterization, we take into consideration only sessions that start and finish within the collection time interval. We characterize traffic volume separately depending on its direction: inbound and outbound. For each analyzed criterion, we determine the statistical distribution that best approximates the measured data, using least-square fit method [20] and visual inspection.

The service request pattern is characterized from traffic logs in terms of the services that are requested by the users within each of their sessions. A service is a request to an application or application class, such as HTTP, e-mail, and P2P, and is usually identified by one or more port numbers where its server answers requests to the service. We use and extend the IANA taxonomy¹ to match ports to services. The extension is necessary because there are some protocols/ports that are well-known, but not registered there, such as port 4662, which is typically used by the peer-to-peer eDonkey protocol [21]. By using the resulting port-to-service translation table, we can transform each user session into a sequence of services.

For each identified user session, we build a Customer Behavior Model Graph (CBMG) [14]. The CBMG is a state transition graph that has one node for each possible service and transitions between these services. A probability is assigned to a transition between two services representing the frequency at which the user requested the services consecutively in the session. The CBMG is a condensed and semantically rich representation of the user behavior since different types of users may be characterized by different CBMGs

¹Internet Assigned Numbers Authority
(<http://www.iana.org/assignments/port-numbers>)

in terms of the transition probabilities. Representative session profiles are identified by clustering the session CBMGs. We employ the k-means algorithm and choose the number of clusters based on the β_{CV} metric, as described in [14].

We further characterize user request patterns in terms of the e-business services accessed within each user session. In other words, the traffic logs are further processed considering only requests to HTTP or HTTPS services, i.e., ports 80 and 443, respectively, typically used by e-business applications. Each distinct IP address requested within a broadband user session (destination IP) is translated into an URL using the `host` Unix command. Finally, URLs sharing the first class domain (i.e., the web site) are grouped together. For instance, `chat.msn.com`, `intl.msn.com` and `help.msn.com` are grouped together into the `msn.com` web site. We then characterize a number of the most popular sites (i.e., most frequently requested domain names), using the following e-business model categorization, proposed in [15]:

Brokerage model : includes web sites which intermediate business-to-business (B2B), business-to-consumer (B2C) and/or consumer-to-consumer (C2C) markets and that charge or receive commission for this service. This category, which includes web sites, such as `ebay.com` and `priceline.com`, can be further broken into eight subgroups, namely marketplace exchange, buy/sell fulfillment, demand collect system, auction broker, transaction broker, distributor, search agent and virtual marketplace.

Advertising model : includes web sites that provide content and services in conjunction with advertising messages in the form of banner ads. Portals, classifieds, user registration, query-based paid placement, contextual advertising and content-targeted advertising are subcategories that fall within this general model. Yahoo, Hotmail and MSN are web sites which operates in this model.

Information Intermediary model : supports web sites whose main goal is to help improve the relationship between buyers and sellers by providing information about both business parts. This e-business web site model includes advertising networks, audience measure services and incentive marketing services. The web sites `akamai.com`, `nielsen-netratings.com` and `doubleclick.com` are examples of this model.

Merchant model : includes web sites, such as `amazon.com` and `barnesandnoble.com`, that provide a means for wholesalers and retailers to sell their goods and services. Subcategories of this e-business model include virtual merchant, catalog merchant, click and mortar, and bit vendor.

Manufacturer Direct model : includes web sites of product makers or services that sell directly to the final consumer. Purchase, lease, license and brand integrated

	<i>Residential</i>	<i>SOHO</i>
Period	12/23/03 - 01/21/04	12/23/03 - 01/21/04
Total # user sessions completed	256,239	61,112
Total # incoming bytes (GB)	11,422	4,135
Total # outgoing bytes (GB)	4,135	1,128
Mean (CV) # sessions completed per user	45 (0.76)	36 (0.74)
Mean (CV) # session duration (hours)	9.80 (5.00)	13.41 (4.29)
Mean (CV) # incoming bytes per session (MB)	46 (5.02)	70 (3.79)
Mean (CV) # outgoing bytes per session (MB)	20 (8.47)	18 (7.89)

Table 1: Summary of the Workloads (CV = Coefficient of Variation.)

content are categories of Web sites which work with this model, such as `dell.com` and `microsoft.com`.

Affiliate model : includes web sites that provide direct links to merchant (partner) web sites. The affiliate web site offers a percentage of their revenue to their partner sites. Subcategories of this model include banner exchange, pay-per-click and revenue sharing.

Community model : includes web sites that are based on user loyalty and that generate revenue with the sale of secondary products and services or voluntary contributions. This model includes web sites involved with open source, public broadcasting and knowledge networks. A typical example is the Open Source Computing web site (i.e., `redhat.com`).

Subscription model : includes content providers, person-to-person networking services, trust services and Internet service providers. Web sites, such as `aol.com`, falling into this major category typically charge the user subscription for a period (day, week, month or year) to provide content and various services.

Utility model : includes web sites that provide services and charge based on the amount of use.

4 Result Analysis

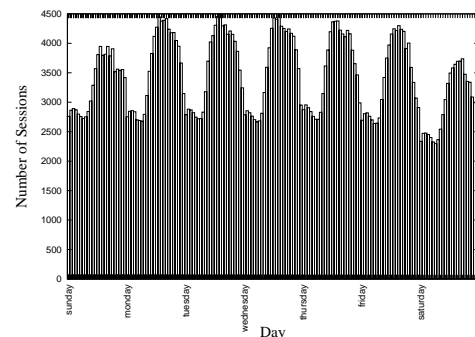
This section shows the most relevant results of our broadband workload characterization. Section 4.1 provides an overview of the residential and SOHO workloads. User session inter-arrival times and durations as well as number of incoming and outgoing bytes transferred within each session in both workloads are characterized in Section 4.2. Section 4.3 analyzes service request patterns and e-business activities.

4.1 Workload Overview

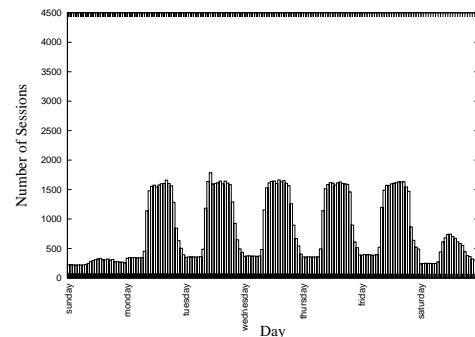
An overview of our residential and SOHO workloads is provided in Table 1. Our logs cover a period of 28 days (12/23/2003 to 01/21/2004), during which a total of over 310 thousand user sessions were completed. More than 80% of them are from residential users. Similarly, over 73% of all incoming bytes and over 78% of all outgoing bytes are from residential users.

An user initiates a session by authenticating himself/herself at the ISP. A session is finished either explicitly by the user or

by timeout after a period of inactivity, which is 4 hours in the case of the ISP that provided the data. Figures 1(a) and 1(b) show the number of simultaneous active (open) sessions during one week for residential and SOHO users, respectively. Figures 2(a) and 2(b) show the same metric over a single day (a Wednesday). Note that most SOHO sessions are active during the day and on weekdays. On the other hand, the fraction of residential sessions active over night and over the weekend is much higher.



(a) Residential



(b) SOHO

Figure 1: Number of Simultaneous Active Sessions (Typical Week).

On average, a residential user completes 1.62 sessions per day and a SOHO user completes only 1.28 sessions per day. This indicates that residential users close their sessions (or are interrupted by timeout) more frequently, during a typical day. On average, a residential user session lasts approximately 9.8 hours, during which 46 MB of data are received and 19 MB of data are sent out. In contrast, typical SOHO user sessions last longer and receive much more data. On average, a SOHO user remains connected for approximately 13 hours, receives

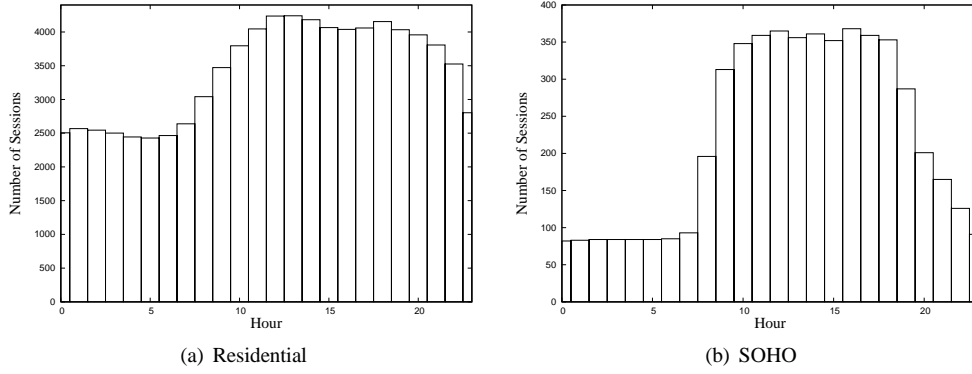


Figure 2: Number of Simultaneous Active Sessions (Weekday).

70 MB of data and sends 18 MB of data. Finally, it is also interesting to note the high variability (i.e., high coefficient of variation) in the number of sessions as well as in the number of bytes transferred within residential and SOHO user sessions. This implies that there might be some heterogeneity among different user sessions within the same category (residential or SOHO).

4.2 Session Characteristics

This section analyzes the first three criteria used in our characterization of broadband user behavior: (i) session arrival process (Section 4.2.1), (ii) session duration (Section 4.2.2) and (iii) the inbound and outbound traffic within a user session (Section 4.2.3).

4.2.1 Session Arrival Process: This section characterizes the user session arrival process during periods of roughly stable session arrival rate in order to avoid spurious effects due to data aggregation. We carefully selected a large number of stable periods covering different times of the day and different days of the week, including weekends.

We found that the user session inter-arrival times are exponentially distributed for both residential and SOHO users, as illustrated in Figures 3(a) and 3(b), respectively, for typical periods of stable arrival rate in each workload. Table 2 summarizes our findings providing the ranges of mean and coefficient of variation (CV) of inter-arrival times as well as the range values of the λ parameter (session arrival rate) of the best-fitted exponential distribution, for all periods analyzed, in each workload. This result is consistent with those presented in [9, 22] for session inter-arrival times.

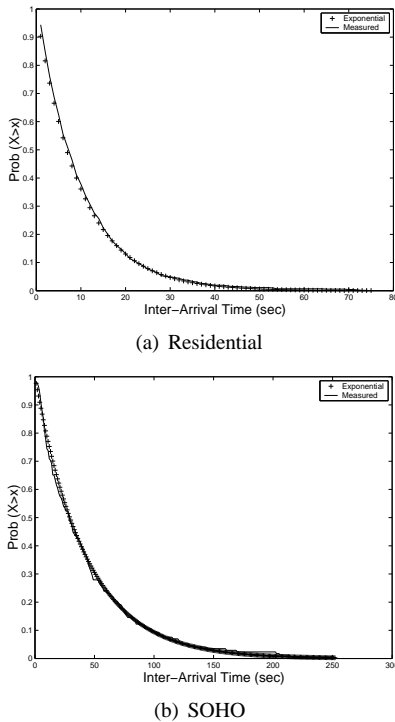


Figure 3: Distribution of Session Inter-Arrival Times (*seconds*).

The tighter range of λ values observed for residential users indicates that residential sessions are initiated at relatively high rates (one each 4 to 10 seconds, on average), throughout the day. In contrast, SOHO users usually initiate their sessions during working hours (as discussed in Section 4.1). In other words, the traditional daily access pattern with peaks in the middle of the day and on weekdays, pointed out [9], is more pronounced among SOHO users.

4.2.2 Session Duration: The durations of residential and SOHO user sessions are characterized separately for different days to avoid data aggregation. For each of the two workloads, we separately characterize the distribution of the durations of all sessions that are initiated on a given day, for a large number of days.

We found that the durations of residential user sessions can be accurately approximated, both at the body and at the tail of the measured data, by a Lognormal distribution, as illustrated in Figure 4(a) for a typical day. This is consistent with results in [9, 22]. In contrast, the duration of SOHO user sessions are better modeled with a combination of a Lognormal distribution, for the body, and a Pareto distribution for the tail. As illustrated in Figure 4(b), the breaking point is around 12 hours.

Workload	Inter-Arrival Times		Exponential
	Mean (sec)	CV	Parameter λ
Residential	4.81 - 10.20	1.02 - 1.05	0.10 - 0.21
SOHO	4.63 - 42.19	0.98 - 0.99	0.02 - 0.22

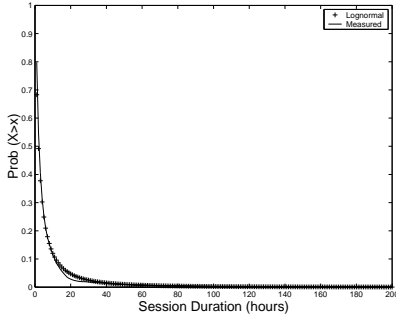
$$\text{Exponential (PDF): } p_X(x) = \lambda e^{-\lambda x}$$

Table 2: Summary of the Distribution of Inter-Arrival Times.

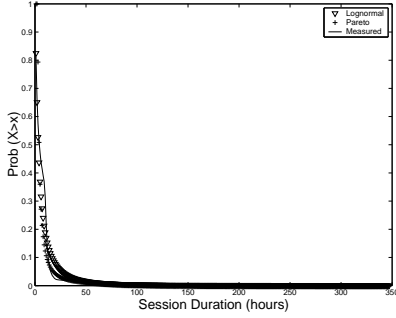
Workload	Mean (hours)	CV	LogNormal Parameters		Pareto Parameters	
			σ	μ	k	α
Residential	4.71 - 13.09	1.75 - 2.47	1.18 - 1.52	0.48 - 1.86	-	-
SOHO	6.95 - 19.21	1.53 - 1.62	0.92 - 1.45	1.04 - 2.30	1.82 - 7.18	1.28 - 1.95

$$\text{Lognormal (PDF): } p_X(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \quad \text{Pareto (PDF): } p_X(x) = \frac{\alpha k^\alpha}{x^{\alpha+1}}, \text{ where } x \geq k.$$

Table 3: Summary of the Distribution of Session Duration (hours).



(a) Residential

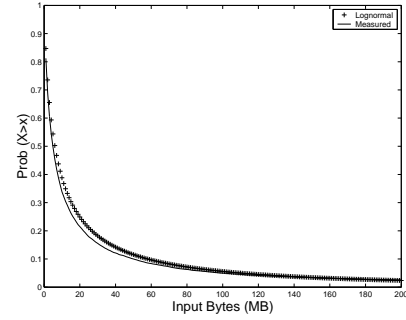


(b) SOHO

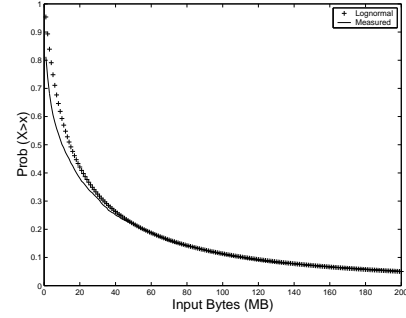
Figure 4: Distribution of Session Duration on a Typical Day (hours).

We speculate this behavior reflects two different classes of SOHO users: (1) users who remain connected mostly while at work, and (2) users who either work longer journeys or remain connected even after leaving the workplace. Table 3 summarizes these results. Note that, on average, a residential user session lasts from 5 to 13 hours. In contrast, an average SOHO user session lasts longer, from 7 to 19 hours.

4.2.3 Inbound and Outbound Traffic: This section characterizes the total numbers of incoming and outgoing bytes, transferred within each user session. As in the previous section, the analysis is performed for different days.



(a) Residential



(b) SOHO

Figure 5: Distribution of the Number of Incoming Bytes per User Session (MB).

We found that, for both residential or SOHO sessions, the number of incoming bytes and the number of outgoing bytes can be each well modeled with Lognormal distributions, as illustrated in Figures 5 and 6, respectively. These results confirm those presented in [4, 6, 9].

Table 4 presents a summary of our results. Compared to residential users, SOHO users typically receive and send larger amounts of data within each session, possibly due to the longer average session duration. Moreover, for each workload, the ratio of the average number of incoming bytes to the average number of outgoing bytes per session is not very high, falling, typically, in the 3 to 5 range. This may be due to the use of services which transfers large amounts of data in

Workload	Metric	Transferred Bytes		Lognormal Parameters	
		Mean(MB)	CV	σ	μ
Residential	Incoming	28 - 44	3.95 - 4.63	1.62 - 1.83	1.76 - 2.46
SOHO	Incoming	47 - 80	3.31 - 3.40	1.47 - 1.70	2.39 - 3.27
Residential	Outgoing	10 - 16	6.82 - 8.27	1.84 - 2.09	0.31 - 1.09
SOHO	Outgoing	9 - 23	2.98 - 6.82	1.51 - 2.09	0.41 - 1.31

Table 4: Summary of the Distributions of the Numbers of Incoming and Outgoing Bytes per Session.

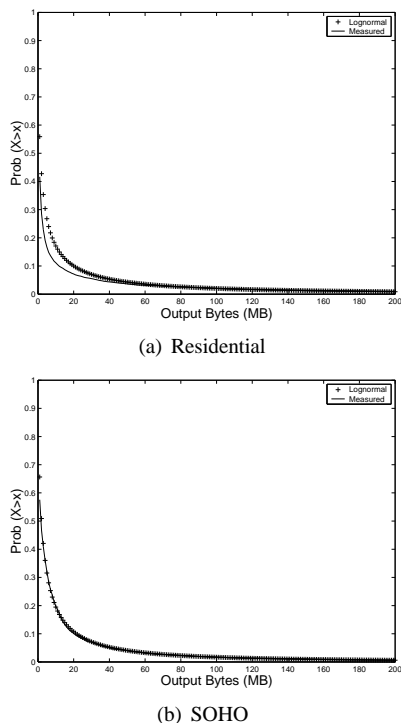


Figure 6: Distribution of the Number of Outgoing Bytes per User Session (MB).

both directions, such as peer-to-peer applications.

In conclusion, compared to SOHO users, residential users usually initiate a larger number of sessions throughout the day. Moreover, residential sessions are typically shorter and transfer fewer bytes, both downstream and upstream.

4.3 User Request Patterns

We now turn to the analysis of the most commonly observed user request patterns within a session in our two broadband workloads. Our analysis focuses on the classes of services (i.e., HTTP, POP3, P2P etc) most commonly requested by the users, and looks further into the HTTP requests in search of the e-business services most frequently accessed by the residential and SOHO broadband users in our workloads.

We first look into the popularity of different services in the SOHO and residential workloads. The popularity of a service is assessed in terms of the percentage of sessions that include at least one request addressed to a port number that identifies the service (e.g., port number 80 for HTTP service).

Figures 7(a) and 7(b) show the popularity of different services among residential and SOHO sessions. Although HTTP appears in over 95% of all sessions in both workloads, e-mail (i.e., POP3, SMTP) as well as interactive applications such as Instant Messenger and ICQ are also popular among both residential and SOHO users. We also point out the significant fraction of sessions that include requests to P2P services, such as Kazaa. In particular, around 23% of the residential sessions and 12% of the SOHO sessions contain requests addressed to P2P services, illustrating the growth of the popularity of these applications among broadband users, previously discussed in [11, 12]. Interestingly, applications that have higher bandwidth requirements such as streaming media are very modestly used by our users.

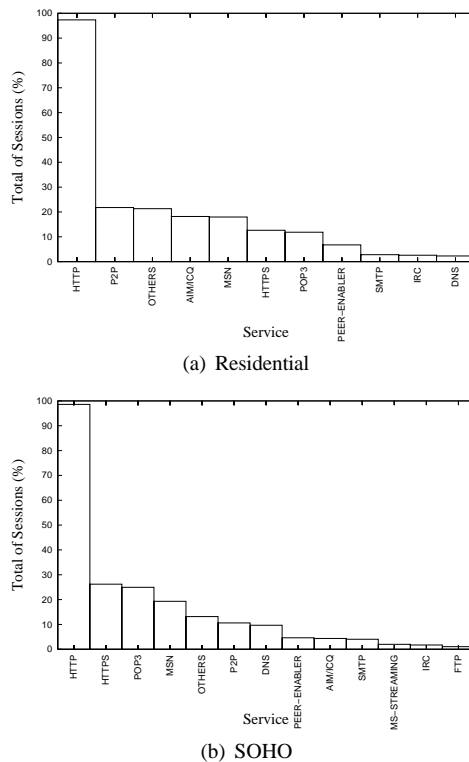


Figure 7: Service Popularity.

The following sections characterize the most common request patterns within a user session in each workload. Section 4.3.1 characterizes the patterns of Internet services requested within a session. Section 4.3.2 focuses on the e-business activities within each user session, characterizing the most frequently requested HTTP-based e-business services.

	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>	<i>Class 4</i>
Services requested within session	HTTP(43%) P2P(24%) MS-DS(19%) EPMAP(14%)	HTTP(97%) HTTPS(3%)	HTTP(73%) MSN(18%) ICQ(6%) POP3(3%)	P2P(68%) HTTP(32%)
Total # sessions (%)	3,249 (4%)	51,606 (66%)	9,797 (13%)	10,640 (14%)
Total # incoming bytes (GB) (%)	214 (7%)	1,436 (47%)	361 (12%)	806 (26%)
Total # outgoing bytes (GB) (%)	75 (6.5%)	451.6(39%)	111 (10%)	405 (35%)
Mean (CV) duration (hours)	9.01 (2.16)	6.06 (2.61)	8.13 (2.48)	10.43 (2.24)
Mean (CV) incoming bytes (MB)	65.88 (3.76)	27.82 (5.09)	36.80 (4.37)	75.79 (3.69)
Mean (CV) outgoing bytes (MB)	23.14 (4.89)	8.75 (10.33)	11.29 (8.18)	38.06 (4.99)

Table 5: Summary of the Main Classes of Residential User Request Patterns.

	<i>Class 1</i>	<i>Class 2</i>	<i>Class 3</i>	<i>Class 4</i>	<i>Class 5</i>
Services requested within sessions	HTTP(77%) POP3(13%) P2P(6%) SMTP(4%)	HTTP(96%) HTTPS(3%) POP3(1%)	HTTP(60%) MSN(34%) POP3(4%)	HTTP(56%) DNS(22%) HTTPS(18%) POP3(4%)	P2P(64%) HTTP(36%)
Total # sessions (%)	1,644 (7%)	16,010 (67%)	2,458 (10%)	1,603 (7%)	1,572 (7%)
Total # incoming bytes (GB) (%)	162.8 (12%)	573.6 (44%)	199.3 (15%)	122.7 (9%)	235.7 (18%)
Total # outgoing bytes (GB) (%)	36.9 (11%)	134.2 (38%)	35.3 (10%)	39.1 (11%)	101.4 (29%)
Mean(CV) duration (hours)	12.23 (2.07)	7.89 (2.53)	11.99(1.50)	11.40 (3.51)	14.82 (1.98)
Mean(CV) incoming bytes (MB)	99.03 (2.60)	35.83 (3.60)	81.08 (2.70)	76.56 (3.26)	149.92 (2.58)
Mean(CV) outgoing bytes (MB)	22.44 (4.81)	8.38 (11.35)	14.37 (3.02)	24.37 (6.85)	64.5 (3.91)

Table 6: Summary of the Main Classes of SOHO User Request Patterns.

4.3.1 Service Request Pattern: The service request pattern is characterized in terms of the frequency of requests to each service and the frequency at which a user switches between different services, within the same session. To do so, we represent the sequence of service requests within each session with a CBMG, as described in Section 3, and use standard clustering techniques to find the most representative per-session service request patterns.

Our analysis uncovered six classes of significantly different request patterns in each workload, summarized in Tables 5 and 6 for residential and the SOHO workloads, respectively. Due to space constraints, we focus only on the classes that accounted for at least 3% of the sessions, omitting two unpopular residential classes and one unpopular SOHO class. Each class is defined by the frequency of requests to each service within a session (first row).

Within both residential and SOHO workloads, the session classes can be further grouped into two major super-classes. One super-class represents those sessions that are dominated by HTTP requests but also may include some requests to other services such as e-mail, Instant Messenger, ICQ and, generally speaking, P2P services. This category consists of classes 1, 2 and 3 in the residential workload, and classes 1, 2, 3 and 4 in the SOHO workload. Compared to sessions consisting mostly of HTTP requests (class 2 in both workloads), the use of e-mail and interactive chatting applications (class 3 in both workloads) increase significantly the average session duration and the average volume of traffic received and sent out. In other words, users remain connected for longer periods communicating with other people. Sessions that include some requests to P2P services (class 1 in both workloads) are even longer and transfer more data, as one might expect.

The second user session super-class is dominated by P2P re-

quests (classes 4 in the residential workload and 5 in the SOHO workload). They last, on average, much longer than the HTTP-based sessions and transfer significantly larger volume of data.

We note the significantly lower coefficients of variation for the session duration and numbers of incoming and outgoing bytes, for each session class, compared to the variations observed in the aggregated SOHO and residential workloads, shown in Table 1. This means that categorizing users based on their request patterns is probably a more effective strategy for QoS analysis and capacity planning than simply relying on its status in the user database.

Figures 8 and 9 show the CBMGs of the request pattern classes identified for residential and SOHO sessions, respectively. As an example, Figure 8(c) shows that a class-3 residential user requesting a HTTP service will switch to a POP3 service with probability 0.03, will switch to ICQ with probability 0.05, will start using Instant Messenger (MSN) with probability 0.12 and, finally, with probability 0.80, he / she will remain requesting HTTP services. Note that, in each class, for both residential and SOHO users, the self-loop transitions have typically high probabilities. In other words, a user tends to use the same kind of service repeatedly.

4.3.2 E-Business Activities: We have shown that the vast majority of user sessions in both residential and SOHO workloads are dominated by HTTP requests. This section further analyzes user requests to HTTP services, focusing specifically on requests to HTTP-based e-business services. As discussed in Section 3, we translate each distinct IP address requested through the HTTP protocol into an URL, group URLs belonging to the same e-business site together, and characterized the popularity of different sites according to the catego-

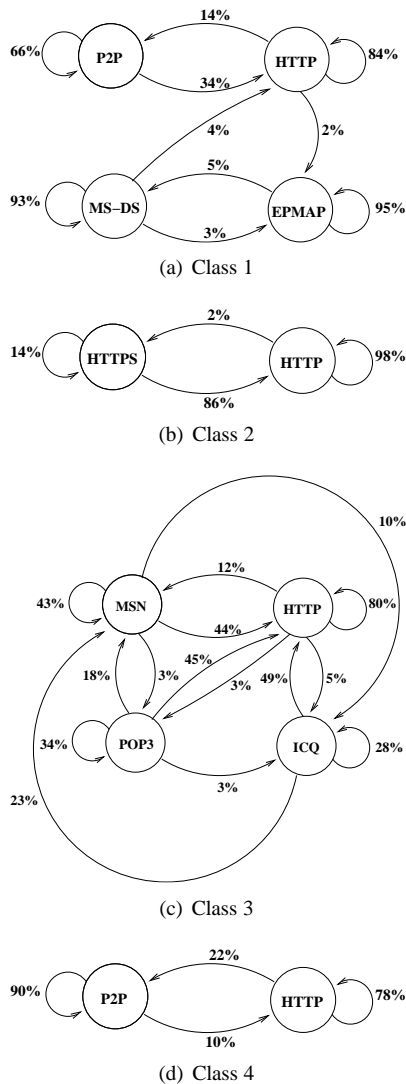


Figure 8: Main Classes of Residential User Request Patterns.

rization proposed in [15].

We found approximately 250 thousand distinct IPs in the two workloads. We were able to translate 78% of them into the corresponding host names using the Unix `host` command. URLs sharing the same third level domain² were grouped into e-business sites. Perhaps surprisingly, we found that the most popular e-business sites among broadband users were mostly the same across all classes of user request pattern (as defined in Section 4.3.1), in both residential and SOHO workloads. Therefore, we first discuss e-business activities considering all user sessions in both workloads. Later, we present some specific findings that are a function of the user class.

We were able to classify 164 out of the 300 most popular e-business sites using the categorization proposed in [15] (36 of them are universities or `.gov` sites and 100 are not reachable from a web browser). These classified e-business sites

²For sake of analysis, we considered domains such as `.com`, `.gov`, and `.org` as second level domains.

account for approximately 60% of all user HTTP requests. Furthermore, the popularity of different e-business sites is highly skewed. For instance, the five most popular classified e-business sites account for 30% of all HTTP requests.

Table 7 shows the percentage of classified e-business sites that fall into each e-business category, as well as the percentage of HTTP requests they account for. Note that no e-business site is categorized as either affiliate, community, or utility. In contrast, e-business sites categorized as either Subscription or Advertising account for 49% of all classified e-business sites and 49% of all user requests to HTTP and HTTPS services, and four of the five most popular classified e-business sites follow the Subscription model, providing text, audio, and video content to subscribed users. More specifically, we noticed that the most popular sites are Brazilian's content services and portals, Yahoo, Hotmail, and Google.

<i>E-business Category</i>	<i>% of classified e-business sites</i>	<i>% all HTTP requests</i>
Subscription	22	26
Advertising	27	23
Information Intermediary	11	7
Manufacturer Direct	16	2
Merchant	23	1
Brokerage	1	1

Table 7: Probability of E-Business Site Categories for All Users.

We also analyzed how the site categories vary across user behavior classes introduced in Section 4.3.1. We observed that the probabilities of the e-business site categories do not vary significantly among user classes. The most popular e-business sites considering the whole user population are also popular in all user classes with slightly variations. The same observation applies to interactive services, such as MSN and ICQ, which present similar popularities for the classes where these services are invoked by users. Finally, there are no differences between the category distribution when contrasting residential and SOHO users, as an indication that the preferences for e-business services are not affected by people being at the office or at home.

5 Conclusions and Future Work

This paper used a quantitative approach for characterizing the behavior of broadband users. The characterization relies on data collected at very specific points in a service provider. The sources of data are the authentication logs, the user database and traffic logs. The characterization was performed at the session level and at the request level and considered residential and SOHO users separately.

Key findings of this study are: (i) both residential and SOHO session inter-arrival times are exponentially distributed, (ii) whereas residential user session duration can be well modeled with a Lognormal distribution, the durations of SOHO

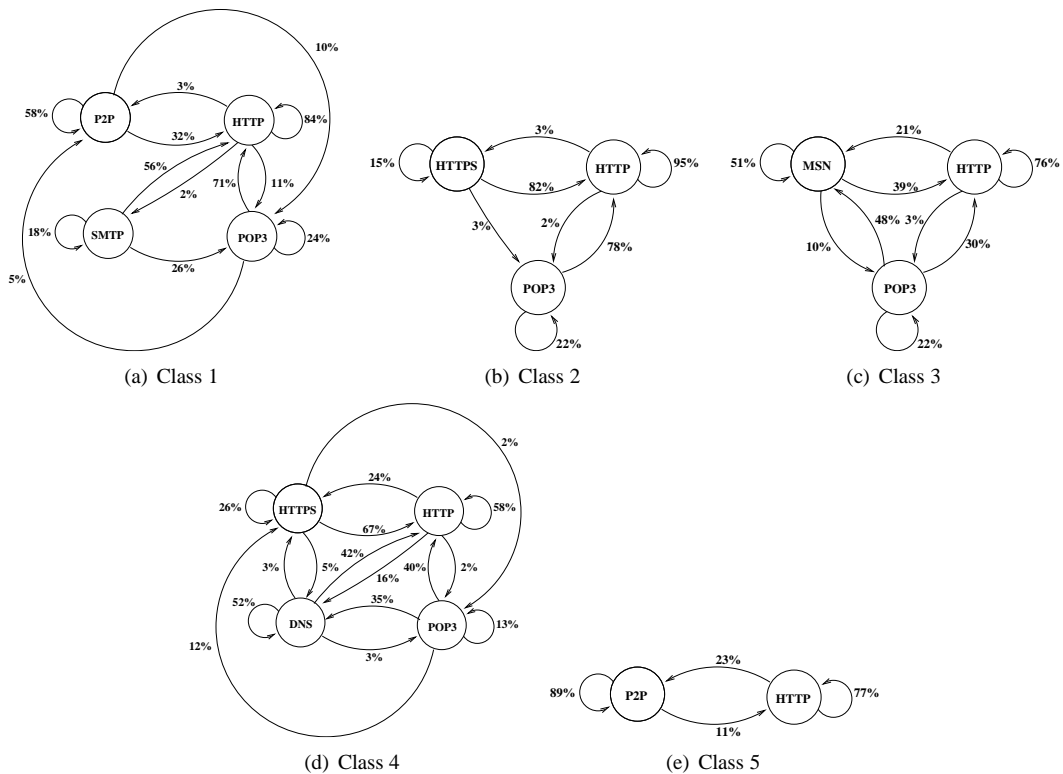


Figure 9: Main Classes of SOHO User Request Patterns.

sessions are better approximated by the combination of a Log-normal and a Pareto distributions, (iii) for both residential and SOHO sessions, the number of incoming and outgoing bytes can be modeled with a Lognormal distribution, (iv) the use of a state transition graph (CBMG) uncovered six classes of significantly different patterns in the user behavior, and (v) user e-business activities concentrate on subscription-based content and services providers and on services that rely strongly on advertisement for generating revenue.

The results presented in this paper are a first attempt to characterize the behavior and e-business activities of broadband users. We are in the process of refining the characterization of the CBMGs, to evaluate the behavior of other services such as games and operating system-oriented services. We are also working to create CBMGs that group classes of users, instead of classes of sessions, and the characterization of broadband daily pattern use.

References

- [1] Netflow. www.cisco.com/warp/public/732/Tech/netflow.
- [2] The broadband difference. Pew & American Life, 2004. www.pewinternet.org.
- [3] M. Arlitt. Characterizing web user sessions. *ACM SIGMETRICS Performance Evaluation Review*, 28(2):50–56, Sep. 2000.
- [4] M. Arlitt, R. Friedrich, and T. Jin. Workload characterization of a web proxy in a cable modem environment. Technical Report HPL-1999-48, Internet Systems and Applications Laboratory - HP Laboratories Palo Alto, Apr. 1999.
- [5] M. Arlitt and T. Jin. Workload characterization of the 1998 world cup web site. *IEEE Network*, 14(3):30–37, May/June 2000.
- [6] P. Barford, A. Bestavros, A. Bradley, and M. E. Crovella. Changes in web client access patterns: characteristics and caching implications. *World Wide Web, Special Issue on Characterization and Performance Evaluation*, 2(1-2):15–28, 1999.
- [7] C. Costa, I. Cunha, A. Borges, C. Ramos, M. Rocha, J. Almeida, and B. Ribeiro-Neto. Analyzing client interactivity in streaming media. In *Proceedings 13th World Wide Web Conference*, New York, NY, May 2004.
- [8] C. R. Cunha, A. Bestavros, and M. E. Crovella. Characteristics of www client-based traces. Technical Report TR-95-010, Department of Computer Science - Boston University, 1995.
- [9] S. Floyd and V. Paxson. Difficulties in simulating the internet. *IEEE/ACM Transactions on Networking*, 9(4), Aug. 2001.
- [10] 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 *Proceedings of the 19th ACM Symposium on Operating Systems Principles (SOSP-19)*, Bolton Landing, NY, Oct. 2003.

- [11] T. Hamada, K. Chujo, T. Chujo, and X. Yang. Peer-to-peer traffic in metro networks: analysis, modeling and policies. *IEEE/IFIP Network Operations & Management Symposium (NOMS 2004)*, Apr. 2004.
- [12] K. Lakshminarayanan and V. Padmanabhan. Some findings on the network performance of broadband hosts. *Internet Measurement Workshop (IMC'03)*, pages 45–50, Oct. 2003.
- [13] N. Leibowitz, M. Ripeanu, and A. Wierzbicki. Deconstructing the kaza network. *3rd IEEE Workshop on Internet Applications (WIAPP'03)*, Jun. 2003.
- [14] D. Menascé and V. Almeida. *Scaling for E-business: technologies, models, performance and capacity planning*. Prentice Hall, Upper Saddle River - NJ, 2000.
- [15] M. A. Rappa. The utility business model and the future of computing services. In *IBM Systems Journal*, 43(1):32–42, 2004.
- [16] C. Rigney, S. Willens, A. Rubens, and W. Simpson. Remote Authentication Dial In User Service (RADIUS). RFC 2865. *IETF*, June 2000.
- [17] C. Rigney. Radius Accounting. RFC 2866. *IETF*, June 2000.
- [18] S. Saroiu, K. Gummadi, R. Dunn, S. Gribble, and H. Levy. An analysis of internet content delivery systems. In *Proceedings of the Fifth Symposium on Operating Systems Design and Implementation (OSDI 2002)*, Dec. 2002.
- [19] S. Sen and J. Wang. Analyzing peer-to-peer traffic across large networks. In *Proceedings of the second ACM SIGCOMM Workshop on Internet Measurement Workshop*, Marseille, France, Nov. 2002.
- [20] K. Trivedi. *Probability & Statistics with Reliability, Queueing, and Computer Science Applications*. 2nd edition, John Wiley & Sons, 2002.
- [21] K. Tutschku. A Measurement-based Traffic Profile of the eDonkey Filesharing Service. In *Proc. of the 5th Passive and Active Measurement Workshop (PAM 2004)*, Antibes Juan-les-Pins, France. April 19-20, 2004.
- [22] E. Veloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. A hierarchical characterization of a live streaming media workload. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, Marseille, France, Nov. 2002.