Characterization and Analysis of User Profiles in Online Video Sharing Systems

Fabrício Benevenuto[1], Adriano Pereira[2], Tiago Rodrigues[1],
Virgílio Almeida[1], Jussara Almeida[1], Marcos Gonçalves[1]

[1] Universidade Federal de Minas Gerais (UFMG)
Department of Computer Science (DCC)
Belo Horizonte, MG, Brazil
`{fabricio, tiagorm, virgilio, jussara, mgoncalv}@dcc.ufmg.br`

[2] Federal Center for Technological Education of Minas Gerais (CEFET-MG)
Department of Computer Engineering (DECOM)
Belo Horizonte, MG, Brazil
`adriano@decom.cefetmg.br`

Abstract. Recently, it has been observed an increasing popularization of video sharing environments. Part of such success is due to the change on the user perspective from content consumer to content creator, basic principle of the Web 2.0. Thus, video service providers are dealing with different challenges, such as content storage, performance and scalability of servers, personalization, and service differentiation. In this context, it is crucial to understand the characteristics of requests that arrive on these servers and the patterns of user navigation on these interactive systems. This work addresses these aspects. Through the analysis of a video service workload from UOL, the largest content provider of Latin America, we present a complete characterization of user sessions, their requests to the server, and their navigation profile. Such analyses are important not only to generate synthetic workload, but also to project and create new infra-structures to video sharing systems. Our results show that there are different users profiles and also provide a better understanding of the user access pattern on video sharing systems.

Categories and Subject Descriptors: H.3.5 [**Online Information Services**]: Web-based services

Keywords: social networks, video sharing systems, video server, Web 2.0

1. INTRODUCTION

Recently, online video sharing systems have been increasing and gaining popularity quickly. Watch and publish video on the Internet is becoming a routine on the daily lives of Web users. According to comScore, in may of 2008, 74% of the North American Internet audience watched videos online, corresponding to 12 billions of videos streamed only in that month [comScore 2008].

Part of this success is associated with the change on the perspective of the user, from simple spectator to an active content creator. Additionally, these environments allow several kinds of interactions between users and videos, such as friendship relations, video evaluation and publication of comments. Associated with this new perspective of the Web, also known as Web 2.0, there are several challenges that the providers of these services need to deal with such as content storage, server performance and scalability, personalization and service differentiation, detection of illegal content, etc. Thus, understanding characteristics of the requests and the patterns of access of users as well as aspects of user navigation when they connect to these sites is important for two main reasons. First, studies of user navigation allow to evaluate the performance of existing systems and lead to better site design [Wilson et al. 2009; Burke et al. 2009] and advertisement placement policies [B. Williamson]. Second, understanding how the workload of social media is re-shaping the Internet traffic is valuable in designing the next-generation Internet infrastructure and content distribution systems [Rodriguez ; Krishnamurthy 2009]. Despite some efforts that characterize user generated workloads [Cha et al. 2007; Gill et al. 2007; Benevenuto et al. 2009], there is not a work that provides a characterization of user navigation from the point of view of a video server.

This work gives the first step in this direction. Through a large data set obtained from the video service of Universo OnLine (UOL)[1], the largest Latin American content provider, we present an in-depth workload characterization of sessions and requests on the video server. Our study uses traces such as clickstream data, which capture *all* activities of users [Chatterjee et al. 2003; Benevenuto et al. 2009]. We obtained a clickstream dataset, which described session-level summaries of over 3.6 million HTTP requests from more than 1 million different IPs during a 26-day period.

Using the clickstream data, we conducted two sets of analyses. First, we characterized the traffic and session patterns of the workload. We examined how frequently people connect to the video server and for how long. Based on the data, we provide best fit models of session inter-arrival times and session length distributions. Second, we provide a definition of user session in our system and we characterize user navigation within sessions. Our analysis unveil dominant user activities and the transition rates between activities.Our study provides many interesting findings, including:

—A typical user session of online video sharing systems lasts about 40 minutes, a high value in comparison with traditional Web systems.
—The popularity distributions of accesses of objects (videos and tags) follow long tails.
—The rankings of user activity in terms of the number of requests sent and sessions created follow long tail and exponential distributions, respectively.
—The arrival request rate at the system presents a periodic pattern with higher intensity during the day and smaller intensity during the night.
—The distributions of inter-request time and inter-session time can be modeled by exponential distributions.
—For longer sessions, users spend more time viewing videos than in short sessions.
—Our analysis reveals different user profiles who access the system, which can be used by system administrators to personalize services.

The remainder of this work is organized as follows. Next section describes related work. Section 3 presents statistics about the workload of the UOL video service. Section 4 discusses the characterization of requests and sessions. In Section 5, we present an analysis of the profile of the users who navigate on the system. Finally, Section 6 concludes the paper and present directions for future work.

2. RELATED WORK

Workload characterization is fundamental to the understanding and improvement of Web systems. There are various studies which present workload characterizations of different types, such as Web servers [Arlitt and Williamson 1996], e-commerce [Menasce and Almeida 2000], blogs [Duarte et al. 2007], video on demand [Costa et al. 2004], and live video [Veloso et al. 2006]. Among various contributions of these works, we highlight the creation of valuable models able to describe the workload that arrives on these server, essential for synthetic workload generation, which allows experimentation and simulation based on realistic workloads. Particularly, Costa *et al.* [Costa et al. 2004] analyzed requests from two video servers in the educational context. They show that the inter-request time

---

[1] http://videolog.uol.com.br

follows a Pareto distribution and the object popularity can be modeled by the concatenation of Zipf-like distributions. Differently, in our work, we present a workload characterization of a user generated content video server. We are not aware of any other work that performs this type of characterization from the point of view of the server.

Complementary to our effort, there are several works that characterize different aspects of online video sharing systems, especially YouTube. In [Cha et al. 2007], the authors analyze the distribution of popularity, evolution, and characteristics of YouTube videos, in addition to evaluate different approaches for the distribution of videos such as caches and P2P sharing. Complementarily, Duarte *et al.* [Duarte et al. 2007] characterizes geographical aspects of interactions of YouTube users. Rodrigues *et al.* [Rodrigues et al. 2010] studied differences on usage statistics and metadata of duplicated videos. Gill *et al.* [Gill et al. 2007] present a workload characterization of YouTube from the point of view of a university, comparing its properties with the Web traffic and other video servers. In [Gill et al. 2008], the authors analyze the characteristics of user sessions on YouTube, by analyzing requests on a university proxy. However, the authors evaluate only aspects such as the session duration and session creation, differently from us, who investigate different actions of users on a session. Zink *et al.* [Zink et al. 2008] perform simulations to show that video caching, on client or in a proxy, and P2P distribution can reduce network traffic and allow a fast access to video in online video sharing systems.

Recently, we presented a comprehensive characterization of the properties of the YouTube *video response network*, that is, the network that emerges from video-based user interactions [Benevenuto et al. 2009]. In [Benevenuto et al. 2009], we further characterize the behavior of three classes of users, namely, legitimate users, spammers and content promoters. Using a machine learning algorithm and exploiting several attributes from the users' profiles, the users' social behavior in the system (i.e., the relationships established among them) and from the user's videos we were able to detect the vast majority of the promoters and spammers. Finally, reference [Benevenuto et al. 2010] provides a comprehensive overview of different sorts of malicious activities in video sharing systems as well as their implications for users and systems.

Differently from these efforts, our work here aims at not only characterizing and understanding the requests that arrive on user generated content video server, but also to investigate and identify the profile of users who access these systems. Complementarily, Benevenuto *et al.* [Benevenuto et al. 2009] used clickstream data to characterized user navigation and social interactions in online social networks, such as Orkut, Hi5, MySpace, and LinkedIn.

3. WORKLOAD DESCRIPTION

In our study, we analyze the workload of the video service of UOL, an important content provider in Brazil and Latin America. The clickstream data obtained correspond to a period of almost one month, from 12/12/2007 a 01/07/2008, accounting for a total of 3,681,232 requests, from more than 1,127,537 different IPs.

Each registry on the workload represents a request sent by a user to the video service. The following information is available for each request: *IP, time, request, status, size, referrer, and agent.* The field *IP* contains the anonymous IP address that generated the request. The field *time* corresponds to the moment, including date and time in seconds, in which the request was received by the server. The field *request* contains not only the URL requested, but also the method and protocol used. The field *status* shows the HTTP protocol response code to the request. The field *size* indicates the size of the request in bytes. The field *referrer* shows the URL from which the visiting request was originated. As an example, if a user on a Web page A visits a link that redirects him/her to a video B, the field *request* contains the URL B and the field *referrer* contains the Web page A. The last field, *agent*, identifies the browser and operating system used.

Group Name	Request Type	Number of Requests	Percentage
1:View	View a video	2,758,883	74.94%
2:User	Video list of a user	218.335	5,93%
	Video list of a user with a certain tag	75,583	2.05%
3:Lists	List of top videos	55,307	1.50%
	List of related videos of a video	32,838	0.89%
4:Interactions	Video evaluation	22,038	0.60%
	Video comment	14,131	0.38%
	Favorite video	10,774	0.29%
5:Search	Search	1,625	0.04%
	List of videos with a certain tag	421,700	11.46%
6:Others	Main page	2,679	0.07%
	Error requests or unformatted registry	67,339	1.82%

Table I. Request Groups

The fields *referrer* and *agent* can be missing in some registries, since users may remove them to increase privacy. Additionally, the field *referrer* can not occur when the user types the URL directly on the browser.

In our workload, there are several types of requests, which we organize into six groups, as shown in Table I. The requests from group 1 correspond to video views. In group 2, we have requests related to list the videos of user and list the videos of a user that contain a certain tag. The third group joins user requests to lists of related videos and lists of top videos. In group 4 we have all requests related to evaluate a video (assign a five-star rating) and user interactions related to include add a video as favorite and post a comment to a video. Group 5 corresponds to requests for content search through the video search engine or through accesses to a tag cloud. Error requests are identified through the field status, according to the definitions presented in [Fielding et al. 1999]. For the analysis of the next sections these requests are not considered. Except by group 6, all groups presented in Table I are used in the analysis of user navigation profile presented in Section 5.

3.1 Limitations

Although our data gives us a unique opportunity to study user activities across video sharing systems, the logs have several limitations. First, we are not able to identify the user IDs in the system. Second, while we have information about the IP addresses of users, this information is anonymized. Thus, we are not able to associate a distinct IP address with each user. This is because many residential ISPs use DHCP to dynamically assign an IP address to each host when it connects to the Internet. When a node disconnects, it releases its assigned IP address, which could then be assigned to a different residential customer. Therefore, we are not able to group multiple sessions into events of a single user. Second, the fields *referrer* and *agent* may be missing in some registries of the log, since users can remove them to preserve privacy. Finally, the field *referrer* may be also missing when the user types the URL directly on the browser.

4. WORKLOAD CHARACTERIZATION

In this section, we present a workload characterization of the UOL video service under different perspectives, modeling various aspects and distributions.

4.1 Session Definition

Before presenting our analyses, we need to define an appropriate session duration for the sessions in our data. A user session is defined as a series of requests performed by a user to a Web site during a certain period of time [Menascé et al. 1999; Arlitt 2000]. In online video social networks, a typical session includes a list of videos by subject, search, video streaming, interaction with other users through the publication of comments and evaluation of videos. These requests are very different from requests on user sessions of the traditional Web sites, which does not provide the same level of interaction between users and objects as occurs in the Web 2.0 systems.

To determine the beginning and the end of a session in the UOL video service it is necessary to analyze the inter-request time between requests from the same user to measure the period of inactivity of that user, since the sessions do not present a registry of login and logout. Thus, it is necessary to perform an analysis to identify the time limit between requests to consider them as belonging to the same session. We consider two consecutive requests as belonging to the same session if the time between them is smaller than this limit, namely *session timeout.*

It is important to choose an appropriate session timeout in order to avoid generating sessions which do not represent the use of the service by users, avoiding to join different moments of the use of the service or to fragment the user navigation. Following the methodology proposed in [Menascé et al. 1999], we evaluate the suitable session timeout for our application.

Figure 1 (left) presents the total number of sessions for different values of session timeout. A value extremely small (e.g. 1 minute) could result in a high volume of sessions. As the value of the session timeout increases, the number of sessions is reduced continuously until stabilizing. The stability occurs around 40 minutes, indicating this value as a suitable session timeout.

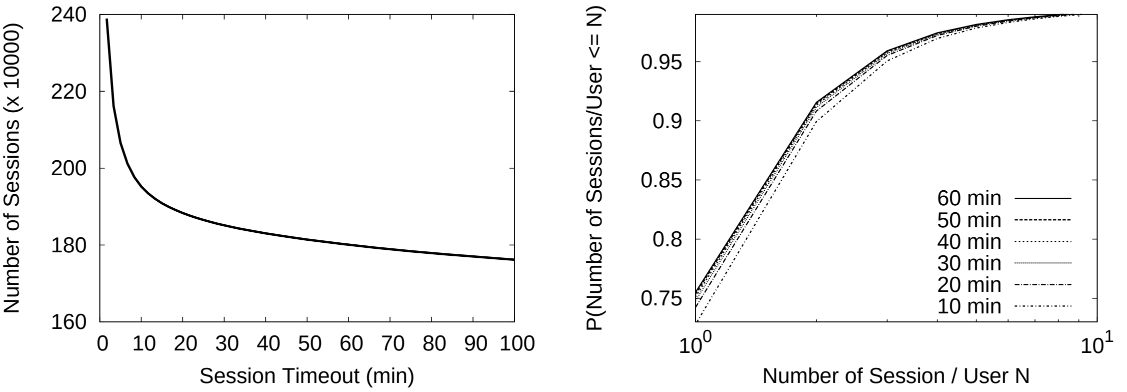


Fig. 1. Session Timeout (left) e CDF of the Number of Session per User (right)

In addition to this analysis to choose a suitable session timeout, we also generate the cumulative distribution function (CDF) of the number of sessions per user for several values of session timeout, as depicted in Figure 1 (right). The difference between the distributions for different values of session timeout is higher for smaller values, becoming very small for values greater than 40 minutes. Thus, we adopt 40 minutes as session timeout for our analyses. In totality, our workload contains 1,127,537 user sessions.

It is interesting to observe that our choice is coherent with the analysis made in [Gill et al. 2008]. Compared to previous efforts, which characterize sessions in traditional Web sites [Arlitt 2000; Oke and Bunt 2002], the longest timeout values obtained are much longer compared to the 10 minutes usually observed. The most intuitive reasons for this behavior are the longer time period that users take to watch a video and the interactive tools, which can make users to spend more time on the site.

4.2 Object Popularity

Initially, we evaluate the popularity of objects aiming at verifying if the popularity of video views and tags searched follow a power law. Power laws establish the following relation $P(E_n) \propto n^{-\alpha}$, where $P(E_n)$ is the probability of reference to the n^{th} most popular element. In order to verify the accuracy of the proposed models, we measure the factor R^2 of the linear regression [Trivedi 2002] for each analyzed distribution. In all the models presented, the values of R^2 are higher than 0.97. A value of R^2 equals to 1 means that there are not differences between the model and the real workload.

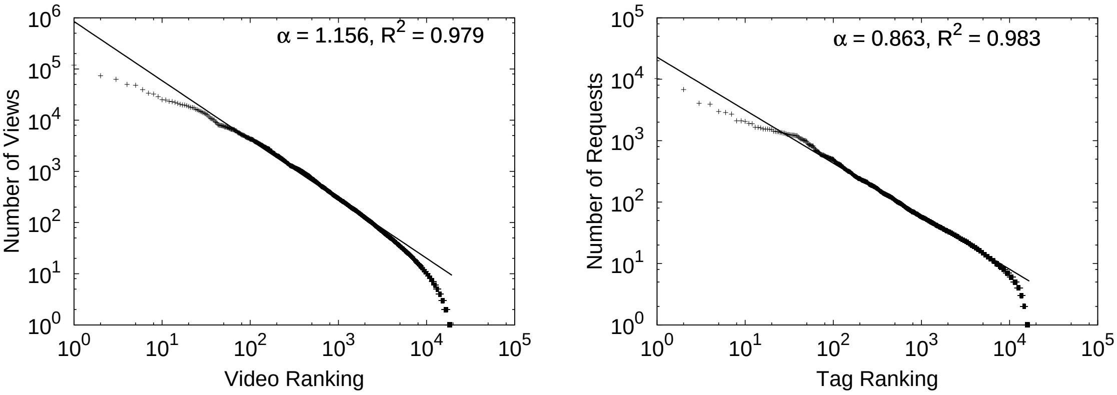


Fig. 2. Videos Ranked by the Number of Views (left) and Tags Ranked by the Number of Access (right)

Next we analyze if the popularity distributions of videos and tags follows a power law. Figure 2 (left) shows the ranking of video sorted by the number of views. We can note that a small number of videos have a high number of views and that a high number of videos have only a few views. Such observation is important since it suggests an opportunity for video caching. In fact, the distribution is modeled by a function that follows a power law, with $\alpha = 1.156$ and $R^2 = 0.979$.

Similarly, Figure 2 (right) shows the ranking of access to tags (e.g. lists of all videos with a certain tag). We can note that some tags concentrate a high number of accesses. As example, the first tag of the ranking has 10,266 accesses. This ranking can be modeled by a power law distribution, with $\alpha = 0.983$ and $R^2 = 0.983$.

4.3 User Activity

Next, we analyze the level of activity of users in the system. We know that users can access the UOL video service several times in the same session or in different sessions. Thus, in order to model the level of activity of users in the system, we characterize the ranking of users in terms of requests sent and in terms of number of sessions created. By user we mean each anonymous IP of our workload.

Figure 3 (left) shows the ranking of users according to the number of requests sent to the server. We can note that there is a small number of users that generate large amounts of requests to the server and a large number of users who sent few requests. In fact, the distribution is well modeled by a power law of the type $f(x) = bx^{-\alpha}$, with $\alpha = 0.745$, and $R^2 = 0.987$.

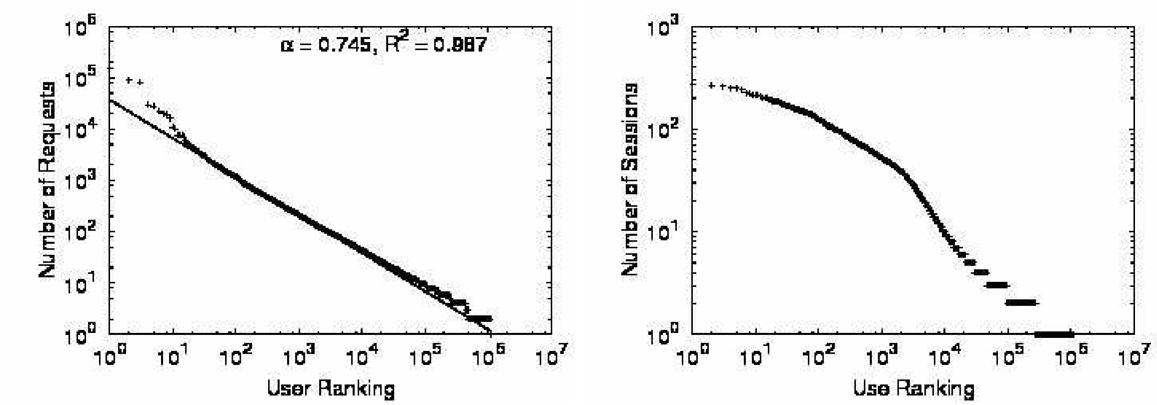


Fig. 3. Ranking of User Activity in Terms of Requests (left) and Sessions (right)

In terms of sessions created in the server, the analysis shows that an exponential distribution is the function that better models the data. Thus, the ranking of sessions can be modeled by an exponential distribution of the type $f(x) = \alpha e^{\beta x}$, with $\alpha = 175.2$ and $\beta = -0.002681$. This result emphasizes the behavior that few users can generate large amounts of sessions, whereas most users generate only a few sessions.

4.4 Temporal Patterns

This Section analyzes the number of requests that arrive in the server as a function of time. The requests about video streaming are not registered in the workload. We have registered only HTML requests that provide access to the video. Thus, we cannot quantify the traffic in terms of bytes transferred by the UOL video service with our workload.

Figure 4 (top left) shows the number of requests that arrive in the server in time intervals of one hour. The curve presents a periodic pattern, with more intensity of requests during the day and small intensity during the night, similarly to other traditional Web servers [Veloso et al. 2006; Arlitt and Williamson 1996]. Note that there are some points in which we can see 50,000 requests in 1 hour. Such points represent links to videos available on popular Web pages of the UOL portal.

In order to analyze the participation of users visiting the system, we characterize the inter-request and inter-session time. We present in Figures 4 (top right) and (down) the complementary cumulative distribution function for these two metrics. We can note that the probability of the inter-request time being higher than 5 seconds is smaller than 1%, whereas 57% of the requests that arrive in the server have time intervals smaller than 1 second. Similarly, about 96% of the intervals between sessions are smaller than 5 seconds.

Both distributions are better modeled by an exponential function of the type $f(x) = \alpha e^{\beta x}$. For the distribution of inter-request times we obtained an $\alpha = 0.424$ and $\beta = -1.298$ with $R^2 = 0.996$, and for the distribution of inter-session times we found an $\alpha = 0.5518$ and $\beta = -0.7309$ with $R^2 = 0.989$.

4.5 Referrer of Requests and Sessions

Next, we analyze the referrer of requests and sessions of users accessing the system. To analyze how users begin to navigate in the video system, we analyze the referrer of the first request of each session. About 50% of the sessions do not have the field referrer in the first request, being thus discarded. Similarly, around 40% of the requests do not have this field and were thus not used. Table II shows the referrer of sessions and of requests accessing the system. We note that most part of the sessions and

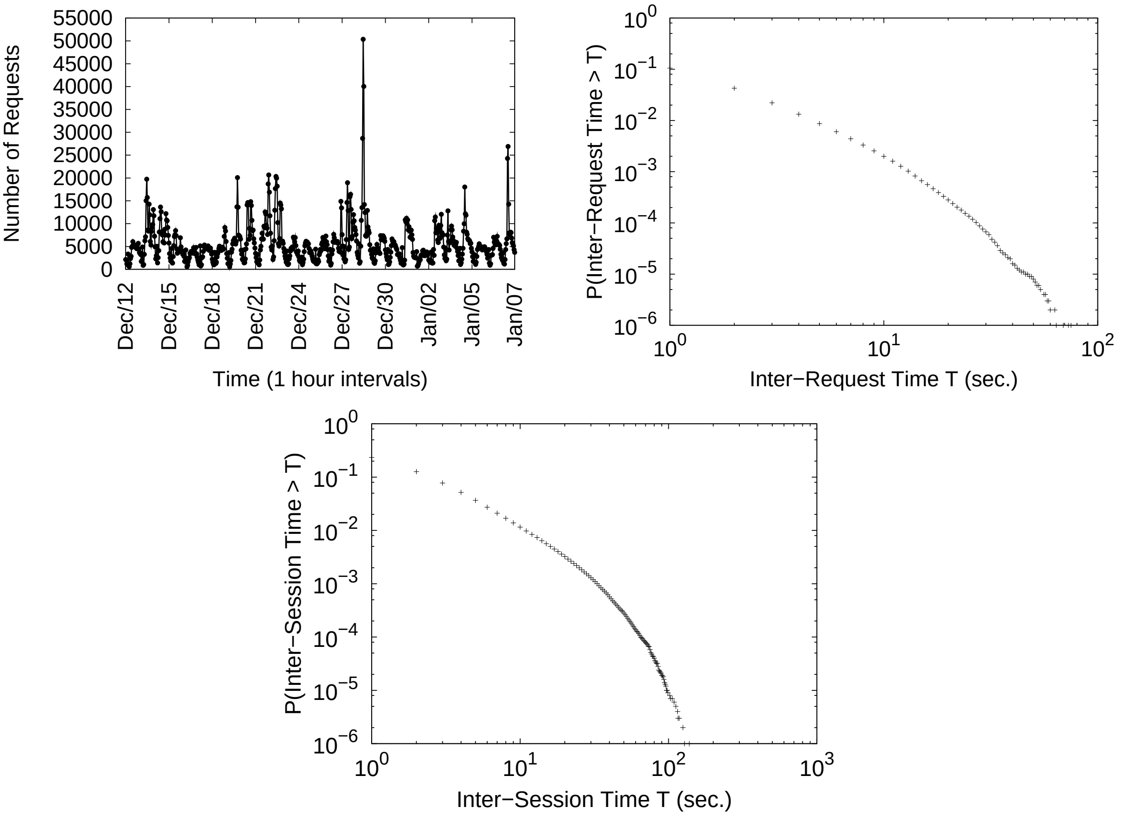


Fig. 4. Number of Requests in Intervals of one hour (top left). CCDF for the Inter-request Time (top right) and Inter-Session Time (down)

Domain	**% of Accesses**	**% of Sessions**
uol.com.br	50.46%	75.71%
videos.uol.com.br	39.58%	12.10%
.br	7.26%	7.89%
Others	2.69%	4.30%

Table II. Referrer of User Requests and Sessions

requests come from other UOL services. However, a significant fraction (around 40%) of the requests have origin in the video service, corresponding thus to users watching other videos or interacting with other users in the system. Only a small part of the requests and sessions come from other sites.

4.6 Probability of Activity over Time

We next investigate whether there is any correlation between the occurrence of a particular type of activity (e.g. Search, view a video, etc.) and session duration. To check for such correlation, we categorized user sessions into four non-overlapping classes based in their session durations: (*a*) less than 1 minute long, (*b*) between 1 and 10 minute long, (*c*) between 10 and 20 minute long, and (*d*) longer than 20 minutes. For sessions belonging to each of these intervals, we examined the average proportion of the total session duration that a user spent on each activity.

Figure 5 shows the fraction of time spent on each type of activity as a function of session duration.

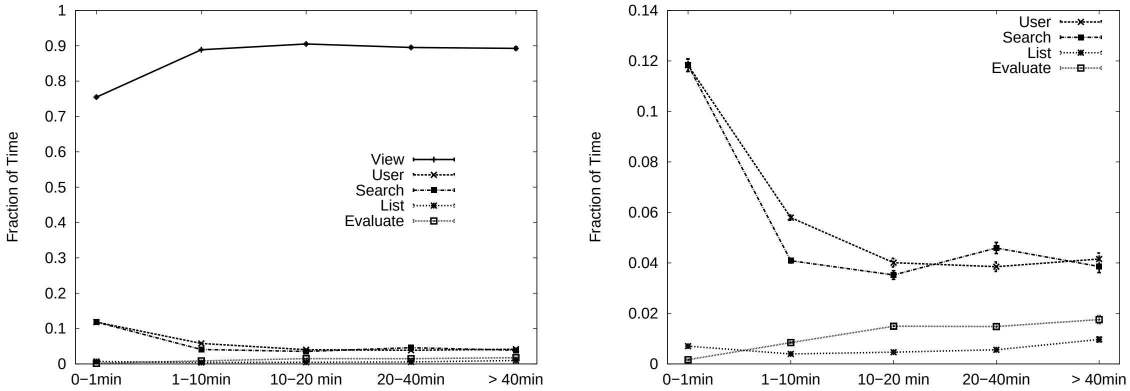


Fig. 5. Probability of Different Types of User Activities as Functions of Session Durations (Error Bars Indicate 95% Confidence Interval)

The results are shown in two separate plots, one containing all groups of activities and the other removing the group View in order to emphasize the trends for the less popular activities. We found two key patterns. First, regardless of session duration, users spent most of their time viewing videos. In very short sessions (i.e., under 1 minute long), users spent 76% of their time in these activities. For longer sessions (i.e., 20 minutes or longer), users spend more than 89% of their session times viewing videos. Second, the remaining categories of activities become less prevalent for longer sessions. The exception is the time spent evaluating content, which increases by a factor of 9 when comparing sessions shorten than 1 minute to those longer than 20 minutes.

5. MODELING USER NAVIGATION PROFILES

This section models user navigation profiles of the UOL video service. In Section 5.1, we present the basic modeling strategy, applying it to build a general model of all users in the system. In Section 5.2, we categorize users into separate groups and analyze the different user navigation profiles.

5.1 Overall User Navigation Profile

In order to understand the navigation profiles of users during their sessions in the system, we build a probabilistic direct graph, where the nodes represent the possible types of user requests (e.g., search, view, etc.) and the arcs represent the navigation between one type of request to another within a single session. Also weights represent probabilities of the navigation pattern occurring. We name this graph as UBMG (User Behavior Model Graph). The UBMG is based on the *Customer Behavior Model Graph* - CBMG [Menasce and Almeida 2000], a methodology to represent the user navigation in e-commerce services. The UBMG nodes correspond to the groups of requests defined in Table I. The *initial* and *final* are introduced to represent the first and the last requests of the user sessions, respectively.

Figure 6 illustrates a typical UBMG, considering all user sessions in our workload. We can note that most users initiate their sessions visualizing videos (86%), whereas the others visit user profiles or perform searches. In constrast, only a tyny fraction of users start their sessions by browsing lists of videos or evaluating videos. After viewing a video, most users tend to keep viewing videos or even to finish the session, although transitions to the other states may occur with non-negligible probability. Interestingly, we find strong self loops in almost all states. For example, search is followed by another search with a probability of 0.562. Similarly, there is a high probability (0.545) of a user keep browsing user profiles repeatedly. Repetition is also evident for browsing lists of videos (probability 0.475). The

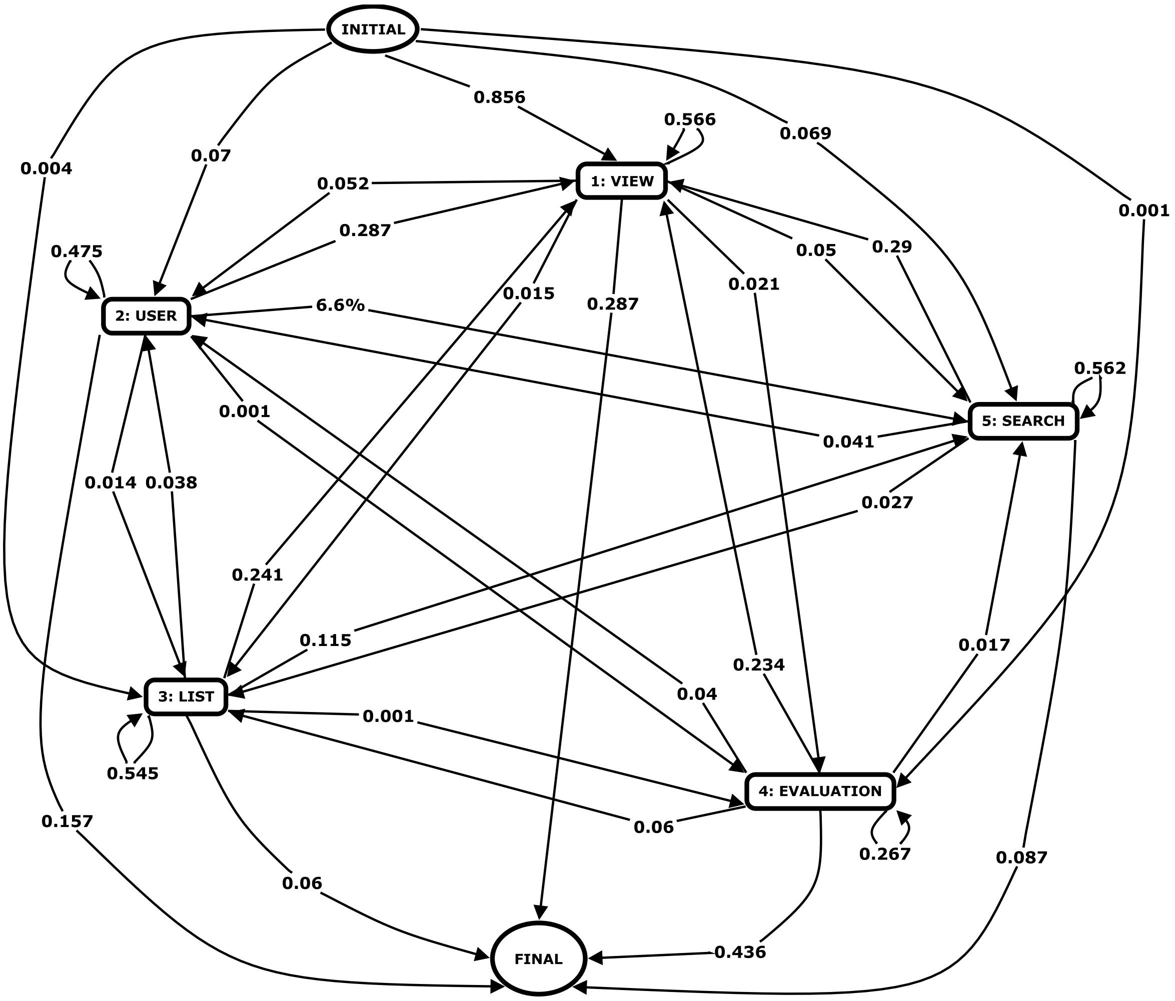


Fig. 6. Typical UBMG: Overall User Behavior

occurence of such high probabilities of repeated activities of the same type, particularly browsing, may help driving the design of prefetching mechanisms.

Next, we propose a method to categorize users into groups according to the different navigation profiles. Such differentiation is important to support the design of customized services.

5.2 Groups of User Navigation Profiles

So far we have examined the overall navigation patterns of users across all sessions. Although the UBMG shown in Figure 6 is useful to uncover the typical navigation pattern in the system, it provides only an overall picture, and thus, does not show the heterogeneity among users in terms of their individual profiles. Next, we propose a method to categorize users into groups according to their navigation profiles.

We start by computing the UBMG of each individual user, considering all her sessions. Next, we apply a clustering algorithm [Bock 2002], in order to identify groups with similar characteristics based on an attribute vector. More specifically, we define each session as an unidimensional vector, where each position in the vector contains the probability of a user navigating from one activity category to another. For each user we compute her individual UBMG across all her sessions and then we use the probabilities of the 35 possible UBMG arcs as user attributes to the clustering algorithm.

We used X-means clustering algorithm [Pelleg and Moore 2000], which extends the popular K-means [Jain et al. 1999] algorithm. A key advantage of X-means over K-means is that the algorithm not only provides the clusters, but also estimates the best possible number of clusters. Therefore, we do not have to decide a priori the number of profiles. X-means algorithm finds clusters by minimizing the sum of the squared distances between each vector and the cluster's centroid, a vector that represents the averaged properties of each group. We consider the Euclidean distance between two vectors, which is computed by as follows:

$$D = \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i)^2 \tag{1}$$

where n is the size of any vector, and x and y are the two vectors.

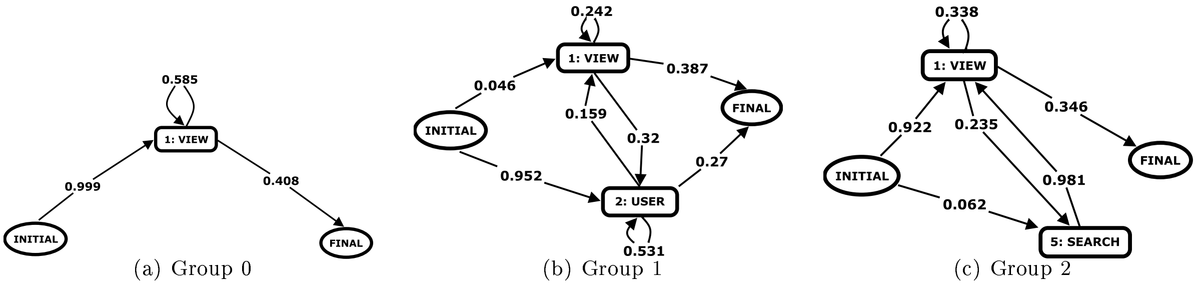
We used the implementation of X-means available on the Weka tool [Witten and Frank 2005] and set the maximum number of groups to 20. The X-means algorithm indicated that 15 distinct groups was the best choice to fit our dataset. Sessions with only one request were discarded as they do not add value to our analysis (i.e., their UBMG representations only have arcs that include the initial or the final state). In total, we discarded 779,384 sessions, focusing our following analysis on the remaining 348,153 sessions and of 345,152 users.

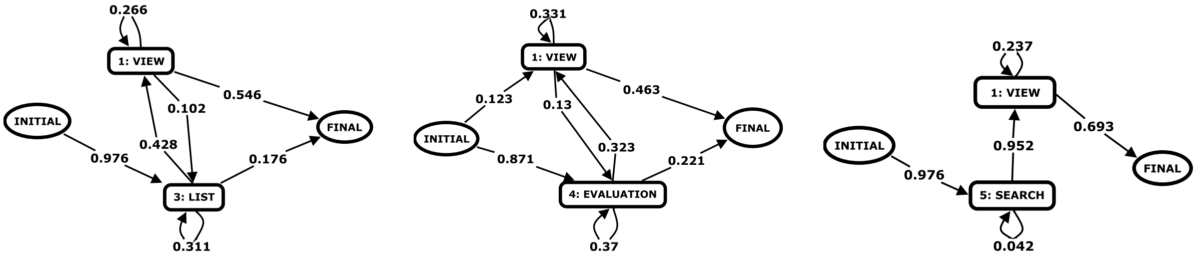
Table III presents the identified groups of identified users, the number number of users, and the frequency of occurrence of each group. It also shows the predominant *Initial* and *Final* transitions of each group, by presenting the state *to* and *from* which a user from each group typically navigates.

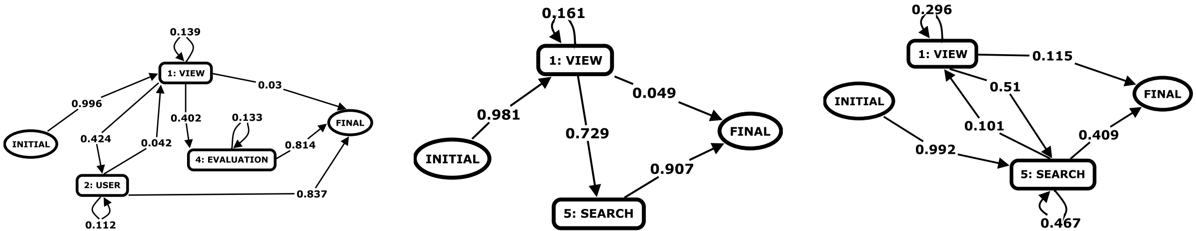
Group	Predominant Transition		Number of Users	Frequency (%)
	Initial *to state*	Final *from state*		
0	1	1	195,028	55.64
1	2	2 e 1	15,102	4.38
2	1	1	11,424	3.31
3	3	1 e 3	1,352	0.39
4	4	1 e 4	273	0.08
5	5	1	13,211	3.83
6	1	2 e 4	28,562	8.28
7	1	5	8,427	2.44
8	5	5	9,296	2.69
9	5	2 e 1	803	0.23
10	2 e 1	3 e 1	366	0.11
11	1	1	33,137	9.60
12	1	1 and others	3,726	1.08
13	1 e 5	1 e 5	6,722	1.95
14	1 e 2	1	20,723	6.00
		Total	345,152	100.00

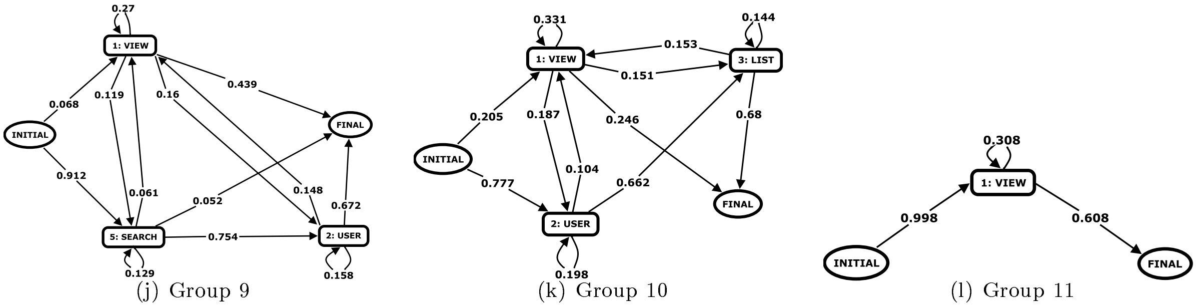
Table III. User Navigation Profile - Groups

We now turn our attention to the UBMGs of all 15 groups, which are shown in Figure 7. We start by discussing the profiles of users who predominantly view videos, here referred to as *Viewers*. These profiles correspond to groups 0, 2, 6, 7, 11, 12, 13, and 14. We note that we omit arcs with probabilities lower than 0.03 for the sake of clarity. We also choose to exclude states having only arcs with low probabilities.


(a) Group 0 (b) Group 1 (c) Group 2


(d) Group 3 (e) Group 4 (f) Group 5


(g) Group 6 (h) Group 7 (i) Group 8


(j) Group 9 (k) Group 10 (l) Group 11

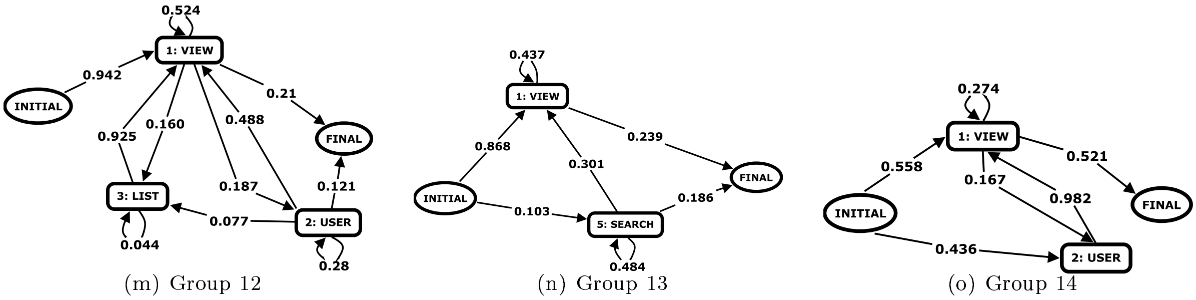

(m) Group 12 (n) Group 13 (o) Group 14

Fig. 7. User Behavior Model Graphs for Different User Profiles

Figure 7(a) presents the graph of the navigation profile of group 0, the most frequent one accounting for 55.64% of the users. These users usually initiate their sessions watching a video. Then 58.5% of them watch other videos, whereas the rest leave the system. Note that, groups 0 and 11, represented on Figure 7(l), are very similar. The difference is that users of group 11 have a small chance of performing other activities (not represented), such as viewing a user profile, evaluating content, and listing videos. Jointly, these two groups represent more than 65% of the users.

Figure 7(g) shows the typical user navigation profile of group 6. Similarly to users from group 0 and 11, these users also initiate their sessions watching a video. However, most of them visit a user profile or evaluate content afterwards. Groups 2 and 7 (Figures 7(c) and 7(h), respectively) are also somewhat similar to groups 0 and 11, whereas group 2 users typically watch videos after searching (probability of 0.981), group 7 users leave the system after the search. One key difference is that after watching at least one video these users may also search for a new video.

The last three groups of *Viewers* are the groups 12, 13, and 14, depicted in Figures 7(m), 7(n), and 7(o), respectively. After viewing a video, group 12 users typically access lists of videos and then view another video or access a user profile, with a high chance (almost 49%) of viewing a video again. Some users from group 13 start their sessions by performing searches (probability of 0.103), whereas a large fraction of users from group 14 start their sessions by browsing a user profile (probability of 0.436) and then, watching videos with probability 0.98.

Next, we analyze profiles of users who start their sessions predominantly by searching for content. These users, referred to as *Searchers*, correspond to groups 5, 8 and 9. Figure 7(f) illustrates the typical profile of group 5. These users start their sessions searching (probability of 0.976) before watching videos with probability 0.95. Similarly, users from group 8 and 9 also start their sessions by searching (probabilities of 0.99 and 0.91, respectively). However, instead of watching videos after searching like most users from group 5, users from group 8 typically keep searching repeatedly, with probability 0.47 and leave the system with probability 0.41. After performing a search group 9 users have a high probability (0.75) to browse a user profile.

Groups 1 and 10, illustrated in Figures 7(b) and 7(k), correspond to users who start their sessions by browsing profiles. Indeed, Users from these groups show very similar profiles with the difference that group 1 users have also some chance (probability of 0.53) of accessing lists of videos maintained by the system.

Only a small fraction (less than 1%) of the users start their sessions by listing videos. These users are represented by group 3 (Figure 7(d)). Basically, they start their sessions listing videos and then they watch videos with probability 0.43 or continue to navigate on the lists of videos.

Lastly, group 4 (Figure 7(e)) exhibits a suspicious profile: some users begin their sessions by *evaluating* videos. This behavior suggests some kind of malicious or opportunistic action, as we would expect that an evaluation should appear only after the user watches at least one video.

6. CONCLUSIONS AND FUTURE WORK

In this work we use a real and representative workload to characterize access patterns of an online video sharing social network and study the user navigation profiles of this system. As results, we provide several statistical models to various system characteristics, such as popularity of videos, users, and tags, inter-request and inter-session time distributions, etc. Our analyses provide new and useful insights about user of online video sharing systems, which may help the design of future synthetic workload generation as well as drive the development of new infra-structures for this kind of service.

We model the navigation patterns of user sessions using the concept of UBMG. Using a clustering technique we provide an analysis of different user profiles accessing the system. Our results can be used to drive service personalization policies as well as content recommendation for users.

As future work, we plan to characterize new workloads of the UOL video service, including aspects of content creation and social interactions. More importantly, we intend to study the aspects that influence the popularity of videos, which are key to an emergent market, the association of advertisements to videos.

Acknowledgements

This work was supported by the Brazilian National Institute of Science and Technology for the Web (CNPq grant no. 573871/2008-6), CNPq, CAPES, Finep, and Fapemig. We also would like to thank Universo OnLine S. A. - UOL (www.uol.com.br) for the data they provided, which made this research possible.

REFERENCES

Arlitt, M. Characterizing web user sessions. *SIGMETRICS Performance Evaluation Review* 28 (2): 50–63, 2000.

Arlitt, M. and Williamson, C. Web server workload characterization: the search for invariants. *SIGMETRICS Performance Evaluation Review* 24 (1): 126–137, 1996.

B. Williamson. Social network marketing: ad spending and usage. *EMarketer Report*, 2007. `http://tinyurl.com/2449xx`. Accessed in March/2010.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. Detecting spammers and content promoters in online video social networks. In *Proceedings of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*. pp. 620–627, 2009.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gonçalves, M., and Ross, K. Video pollution on the web. *First Monday* 15 (4, April, 2010.

Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Ross, K. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)* 5 (4): 1–25, 2009.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. Characterizing user behavior in online social networks. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC)*. pp. 49–62, 2009.

Bock, H. *Data mining tasks and methods: Classification: the goal of classification.* Oxford University Press, Inc., New York, NY, USA, 2002.

Burke, M., Marlow, C., and Lento, T. Feed me: Motivating newcomer contribution in social network sites. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*. pp. 945–954, 2009.

Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*. pp. 1–14, 2007.

Chatterjee, P., Hoffman, D. L., and Novak, T. P. Modeling the clickstream: implications for web-based advertising efforts. *Marketing Science* 22 (4): 520–541, 2003.

comScore, R. Americans viewed 12 billion videos online in may 2008. http://www.comscore.com/press/release.asp?press=2324, 2008.

Costa, C., Cunha, I., Vieira, A., Ramos, C., Rocha, M., Almeida, J., and Ribeiro-Neto, B. Analyzing client interactivity in streaming media. In *Proceedings of the World Wide Web Conference (WWW)*, 2004.

Duarte, F., Benevenuto, F., Almeida, V., and Almeida, J. Geographical characterization of YouTube: a latin american view. In *Proceedings of the Latin American Web Congress (LAWEB)*. pp. 13–21, 2007.

Duarte, F., Mattos, B., Bestavros, A., Almeida, V., and Almeida, J. Traffic characteristics and communication patterns in blogosphere. In *Proceedings of the Conference on Weblogs and Social Media (ICWSM)*, 2007.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. *RFC 2616: Hypertext Transfer Protocol – HTTP/1.1.* The Internet Society, 1999.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. YouTube traffic characterization: A view from the edge. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*. pp. 15–28, 2007.

Gill, P., Arlitt, M., Li, Z., and Mahanti, A. Characterizing user sessions on YouTube. In *IEEE Multimedia Computing and Networking (MMCN)*, 2008.

Jain, A., Murty, M., and Flynn, P. Data clustering: a review. *ACM Computing Surveys* 31 (3): 264–323, 1999.

Krishnamurthy, B. A measure of online social networks. In *Conference on Communication Systems and Networks (COMSNETS)*, 2009.

Menasce, D. and Almeida, V. *Scaling for E Business: Technologies, Models, Performance, and Capacity Planning.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000.

Menascé, D., Almeida, V., Fonseca, R., and Mendes, M. A methodology for workload characterization of e-commerce sites. In *Proceedings of the ACM conference on Electronic commerce (EC)*, 1999.

Oke, A. and Bunt, R. Hierarchical workload characterization for a busy web server. In *Proceedings of the Int'l Conference on Computer Performance Evaluation, Modelling Techniques and Tools (TOOLS)*, 2002.

Pelleg, D. and Moore, A. X-means: Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Int'l Conf. on Machine Learning (ICML)*, 2000.

Rodrigues, T., Benevenuto, F., Almeida, V., Almeida, J., and Gonçalves, M. Equal but different: A contextual analysis of duplicated videos on youtube. *Springer Journal of the Brazilian Computer Society* 16 (3): 201–214, 2010.

Rodriguez, P. Web infrastructure for the 21st century. *WWW'09 Keynote, 2009.* `http://tinyurl.com/mmmaa7`. Accessed in March/2010.

Trivedi, K. S. *Probability and statistics with reliability, queuing and computer science applications.* John Wiley and Sons Ltd., Chichester, UK, 2002.

Veloso, E., Almeida, V., Jr., W. M., Bestavros, A., and Jin, S. A hierarchical characterization of a live streaming media workload. *IEEE/ACM Transactions on Network (TON)* 14 (1): 217–230, 2006.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. User interactions in social networks and their implications. In *Proceedings of the ACM European Professional Society on Computer Systems (EuroSys)*. pp. 205–218, 2009.

Witten, I. and Frank, E. *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

Zink, M., Suh, K., Gu, Y., and Kurose, J. Watch global, cache local: YouTube network traces at a campus network - measurements and implications. In *Proceedings of the IEEE Multimedia Computing and Networking (MMCN)*, 2008.