

Hierarchical Characterization and Generation of Blogosphere Workloads

Fernando Duarte, Bernardo Mattos, Jussara Almeida and Virgilio Almeida ^{*,1}

Computer Science Department, Federal University of Minas Gerais, Belo Horizonte, MG 30161, Brazil

Mariela Curiel

Computer Science Department, Simón Bolívar University, Caracas, Apdo. 89000, Venezuela

Azer Bestavros ²

Computer Science Department, Boston University, Boston, MA02215, USA

Abstract

We present a thorough characterization of the access patterns in blogspace, which comprises a rich interconnected web of blog postings and comments by an increasingly prominent user community that collectively define what has become known as the blogosphere. Our characterization of over 35 million read, write, and management requests spanning a 28-day period is done at three different levels. The *user view* characterizes how individual users interact with blogosphere objects (blogs); the *object view* characterizes how individual blogs are accessed; the *server view* characterizes the aggregate access patterns of all users to all blogs. The more-interactive nature of the blogosphere leads to interesting traffic and communication patterns, which are different from those observed for traditional web content. We identify and characterize novel features of the blogosphere workload, and we show the similarities and differences between typical web server workloads and blogosphere server workloads. Finally, based on our main characterization results, we build a new synthetic blogosphere workload generator called GBLOT, which aims at mimicking closely a stream of requests originating from a population of blog users. Given the increasing share of blogspace traffic, realistic workload models and tools are important for capacity planning and traffic engineering purposes.

Key words: Workload Characterization, Blogosphere, Workload Generation, Performance

1 Introduction

A distinct and rapidly growing segment of web content available on the Internet is the content in “blogspace” – an interconnected web of what could be best described as a *web log* (blog) of news, opinions, and commentaries maintained by an individual (the blog author, or blogger). While most blogs are related to a subject of general interest (*e.g.*, politics, sports, and technology, *etc.*), many blogs have a more specific or target audiences (*e.g.*, personal diaries, instructor notes for a class or course, *etc.*). As with regular web pages, a typical blog combines textual content with multimedia content, and incorporates links to other blogs, blog entries, and web pages.

An important feature of most blogs is the ability of their readers to leave comments (moderated or not), which themselves become an integrated part of the blog, may elicit further comments by other blog readers, and may trigger the addition of new entries in the same blog or in other blogs. As such, blogs are in fact snapshots of an interactive (“live”) exchange between the various players in blogspace. Such exchanges could be either within a single blog (intra-blog exchanges and references) or across blogs (inter-blog exchanges and references).

A unique characteristic of blogs relates to how their contents evolve over time. Unlike traditional web pages that are mostly static, undergoing arbitrary content modifications over time (including content deletion or substitution), which are hard to trace over time [1], most blog contents change in a very prescribed fashion – namely by having new entries or comments appended to a blog. Thus, by and large, the overall content of a blog (*i.e.*, not just what is rendered on the front page) is monotonically increasing over time. Moreover, blog threads are timestamped (and typically maintained in a reverse-chronological order), and hence clearly serializable. In many ways, the blogspace could be considered as the web-counterpart of Usenet newsgroups, with the exception that blog “ownership” is explicit, whereas for newsgroups, it is more or less collective.

* Corresponding author.

Email addresses: {fernando,bemattos,jussara,virgilio}@dcc.ufmg.br (Fernando Duarte, Bernardo Mattos, Jussara Almeida and Virgilio Almeida), mcuriel@ldc.usb.ve (Mariela Curiel), best@cs.bu.edu (Azer Bestavros).

¹ F. Duarte, B. Mattos and V. Almeida are partially supported in part by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 20060520221328a.

² Supported in part by NSF awards #072064, #0735974, #0524477, #0520166, and #0205294.

1.1 Motivation

Given the prominence and continued growth of blogspace, it is natural to ask whether its characteristics are similar to those of more traditional segments of web content. Indeed, over the last few years, there has been a number of studies that explored the various aspects of the blogspace. For example, the works in [2–4] explored the overall scope, structure, and bursty growth patterns of the blogspace and the social networks underlying them. Such studies are important because they allow us to predict the impact of blogs on various applications, such as web search and social network mining, for example.

Another important consideration relates to the patterns of access targeting the blogspace – and in particular how such *access patterns* impact the portion of *web traffic* induced by the blogspace. Studies that focused on the access patterns for traditional web content have uncovered important properties that were crucial in explaining observed traffic characteristics [5], and were instrumental in building workload models and in developing synthetic traffic generation tools [6]. In this paper we focus on this dimension of blogspace characterization – a dimension that emphasizes impact on traffic and communication patterns, as opposed to the characteristics of higher semantic levels (such as information diffusion through blogspace [7] or the evolution of the topological structure of blogspace [8]).

1.2 Basic Definitions

Throughout this paper, we use the term *blogspace* to refer to the subset of content on the web (*a.k.a.* web pages) that are organized as web logs, or *blogs*. We use the term *blogosphere* to refer to blogspace *and* the community of users (and underlying social networks) accessing it. We use the term *blogger* to refer to the author (or owner) of a blog, an individual who keeps, updates, and otherwise manages the blog. We use the term *visitor* to refer to any user who accesses a blog. We use the term *posting* to refer to an entry created by a blogger on his/her blog, and we use the term *comment* to refer to feedback or comment written by a visitor in response to a specific posting or comment thereof. We use the term *request* to refer to an access (get or post) to a blogosphere server. We use the term *session* to refer to a sequence of consecutive accesses by a single visitor to a blog in a short time span. A *blurker* is a blog reader who does not post comments, and a *commenter* is someone who leaves remarks/comments [9]. When necessary, we differentiate between a *write session* and a *read session* based on whether, in addition to accessing the blog, the visitor submitted a comment or not. The term *session* will be used for sessions that can have both types of requests.

1.3 Hierarchical Characterization Approach

In this paper we adopt a hierarchical approach to characterize the blogosphere workload [10]. Our characterization is accomplished at three levels: user, object and, at the bottom, server level. Blogosphere users are visitors, who can be bloggers as well. During a session, a user makes requests on blogs (objects): posts, reads and comments. Such operations generate HTTP requests to the server. Requests from different blogosphere users arrive at the server as an aggregated workload. This paper focuses on the analysis and characterization of traffic characteristics and communication patterns in these three levels.

1.4 Goals and Contributions

Based on an extensive real workload from a major Internet Service Provider – a workload that consists of over 32 million requests to over 210,000 blogs that resulted in almost one TeraByte of transferred content over a 4-week period – we provide a statistical analysis of how visitors read blogs and make comments in blogosphere, and how bloggers update their blogs. Our study looks at blog accesses as defining blogosphere dialogues. As we show in our work, such interactions range from one-way interactions (from author to readers) to multi-way iterative interactions (dialogue among readership). This more-interactive attribute of blogosphere access patterns leads to interesting traffic and communication patterns, which are different from those observed in access patterns for traditional web content. In that respect, we identify and characterize novel features of the blogosphere workload and discuss the similarities and differences between typical web workloads and blogosphere workloads.

Given the increasing share of Web traffic, synthetic traffic generation is an important tool for capacity planning and traffic engineering purposes. For this reason we used some of our main characterization results to design and implement GBLOT, a synthetic blogosphere workload generator, which can be used to evaluate the short-term behavior of blog hosting sites. The GBLOT architecture has three main components: i) distributional data, ii) a master process and iii) user processes. The master process instantiates the user processes based on blog access patterns and also controls the experimentation time. User processes emulate user behavior based on behavior model graphs. Blog accesses are generated according to the distributional data that characterize the blogosphere workload. The workload generated by GBLOT shows a number of properties that are quite similar to the ones we found in the logs from real blog servers.

1.5 Paper Outline

The remainder of this paper is organized as follows. In Section 2, we give a high-level description of the data sets used in our blogosphere characterization. This is done hierarchically. The top level characterizes how individual users interact with the blogosphere. This is the blogosphere *user view* of the workload, which we present in Section 3. Next, the characterization of how individual blogosphere objects (blogs) are accessed is presented. This is the blogosphere *object view* of the workload, which we present in Section 4. Finally, in the bottom level we characterize the aggregate access patterns of all users to all blogs. This is the blogosphere *server view* of the workload, which is presented in Section 5. The user behavior model implemented by GBLOT is described in Section 6. The design of GBLOT and its validation is described in Section 7. We put our work in context by reviewing related research in Section 8, and we conclude with a summary of our findings in Section 9.

2 Blogosphere Workload Description

In this paper we consider the blogosphere spanned by three anonymized traces from a highly popular Brazilian weblog service. The first trace, which we call the *read-trace*, contains all the read requests to the content of the blogs. The second trace, which we call the *write-trace*, contains all comments sent by users. The third trace, which we call the *admin-trace*, contains all the administrative/management activities on the blogs by their owners, *i.e.*, blog creation, edition, deletion, saving, publishing and posting.

2.1 Trace Format

Each entry in the traces refers to a blogosphere access described using the following syntax:

```
hostname date request status size referrer agent
```

The `hostname` is the IP address which generated the request (whether get or post). The `date` field indicates the day and time the request was made. In the read-log, the `request` field refers to the object requested by the user for reading. In the write-log, the `request` field refers to the comment (and associated blog and posting) written by the user. In the admin-log, the `request` field refers to the object (a blog and posting) that the blogger is manipulating. The `status` field provides the HTTP response code for that request. The

`size` field indicates the size of the data in bytes sent back to the client in response to the request. The `referrer` field indicates the URL of the web page or blog from which the visitor performed its access. The `agent` field identifies the browser and platform used to make the request.

2.2 Trace Sanitization

The requests recorded in the access logs reflect those made by “real” users as well as those made by crawlers of search engines and webbots. Search engines usually identify their crawlers using the `agent` field (*e.g.*, using “Googlebot” and “Yahoo! Slurp” to identify the Google and Yahoo! crawlers, respectively). Since crawlers are not real actors, and hence do not underscore social relationships in the blogosphere, we have identified and isolated all such requests, which amounted to 13,622,219 requests across all three traces. We also eliminated a total of 4,289,007 requests that resulted in redirections (`status` code 301 or 302) or errors (`status` codes 4xx) across all three traces. The analysis presented in this paper excludes all such requests.

2.3 Summary Statistics

As evident from the summary shown in Table 1, the blogosphere encompassed by our traces is sizeable. It consists of over 32 million blog (read) requests and about 278 thousand comment (write) requests. These requests were made by over 4M visitors over a period of four weeks extending from January 12th to February 9th, 2006. During this period of time a total of over 992 GB of data was transferred, over 210K distinct blogs were requested, and over 81K postings to over 30K blogs received at least one comment.

Here we note that our workload (and our characterization thereof) contains all comments submitted by visitors, including those which were deleted or not authorized by bloggers to appear on their blogs. Therefore, we argue that our analysis of comments (especially as it relates to popularity of blogs, postings, and levels of interactions) is more representative than an analysis based on results collected by crawlers to characterize the distribution of responses to blog postings.

In terms of the type of objects requested and served within our blogosphere, our analysis of the traces revealed an almost 2-to-1 split between requests to rendered content and code (62% of all requests were to HTML-type objects, whereas 36% of all requests were to Javascript-type objects), with HTML objects constituting the bulk (97%) of the total bytes transferred.

Table 1

Summary statistics of the blogosphere traces used in this paper (excluding requests by crawlers and requests that resulted in redirections or errors).

Access Log Characteristics	Value
Trace duration	28 days
Trace start date	01/12/2006
Total bytes transferred in GB	992.79
Number of visitors	4,193,371
Number of read requests	32,369,178
Number of write requests (comments)	277,709
Number of admin. sessions	250,271
Number of admin. requests	967,220
Number of blogs in read-log	210,738
Number of blogs in admin-log	74,405
Number of blogs in write-log	30,145
Number of commented postings in write-log	81,561

3 Blogosphere User View

In this section, we focus on the *user view* of the blogosphere workload – how individual users access the blogosphere. To do so, and rather than viewing our traces as sequences of individual requests (reads and writes), we group such requests in sessions. We define a *user session* as the interval of time (and the set of requests within that interval) during which a single user (identified by unique values of the `hostname` and `agent` fields) is “actively” engaged in accessing the blogosphere. A session starts with the first request by the user and ends when the user navigates to other web pages out of the blogspace or when the time since the last request in the session exceeds a timeout value (which we take to be 30 minutes).

Over the entire trace, each user (re)visits the blogosphere any number of times, indicating some level of “interest” in the content. To characterize the interest profile of the blogosphere user population, Figure 1 shows the frequency of user accesses (read requests and comments) versus the interest rank of the user, where the i^{th} ranked user is the one issuing the i^{th} -most requests to the blogosphere. The relationship in Figure 1 underscores a power law with $\alpha = 0.83$ (for read requests) and $\alpha = 0.54$ (for comments), both with $R^2 = 0.99$.

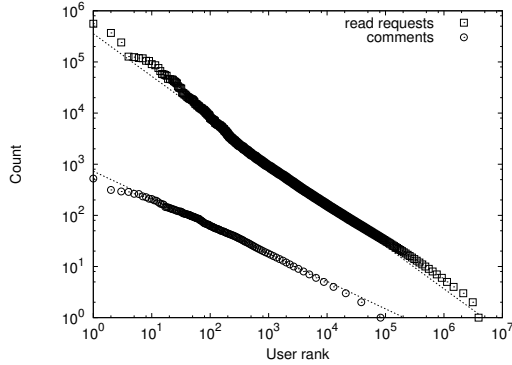


Fig. 1. User access frequency versus rank of user interest.

4 Blogosphere Object View

In this section, we focus on the second level of the hierarchy: the *object view* of the blogosphere workload – how individual objects (blogs) are accessed.

4.1 Popularity Profile

It has been well established that typical web workloads exhibit a significant skew in terms of the popularity of the various objects (web pages) accessed in these workloads [11]. Thus, a natural question with respect to objects in blogspace (namely, blogs) is whether they exhibit a similar popularity profile. Figure 2 shows the popularity profile of the blogs in our blogosphere. Figure 2 (top left) plots, on log-log scales, the number of accesses (read and write) to a blog against the rank of the blog. Approximately 90% of all read requests and 60% of all posted comments target only 10% of all blogs. The skew is even more pronounced if one looks at the most popular blogs: 21 blogs (0.01% of all blogs) account for 7.5 millions read requests (23% of all read requests) in the workload.

In fact, Figures 2 show that the popularity of objects in blogspace, expressed by different popularity estimates, follows a general power law with skew parameter α . Using the total number of read requests to a blog as indicative of popularity yields a skew of $\alpha = 0.97$ (coefficient of determination $R^2 = 0.96$ [12]). Using the total number of posted comments to a blog as indicative of its popularity yields a smaller skew of $\alpha = 0.70$ ($R^2 = 0.97$). Figure 2 (top right) shows that the same skewed popularity profile holds if one considers the number of postings with at least one comment in each as a measure of blog popularity. Similarly, Figure 2 (bottom) shows that the same skewed popularity profile holds if one considers the total number of sessions, the number of write sessions, or the number of distinct users that accessed the blog as the measure of blog popularity.

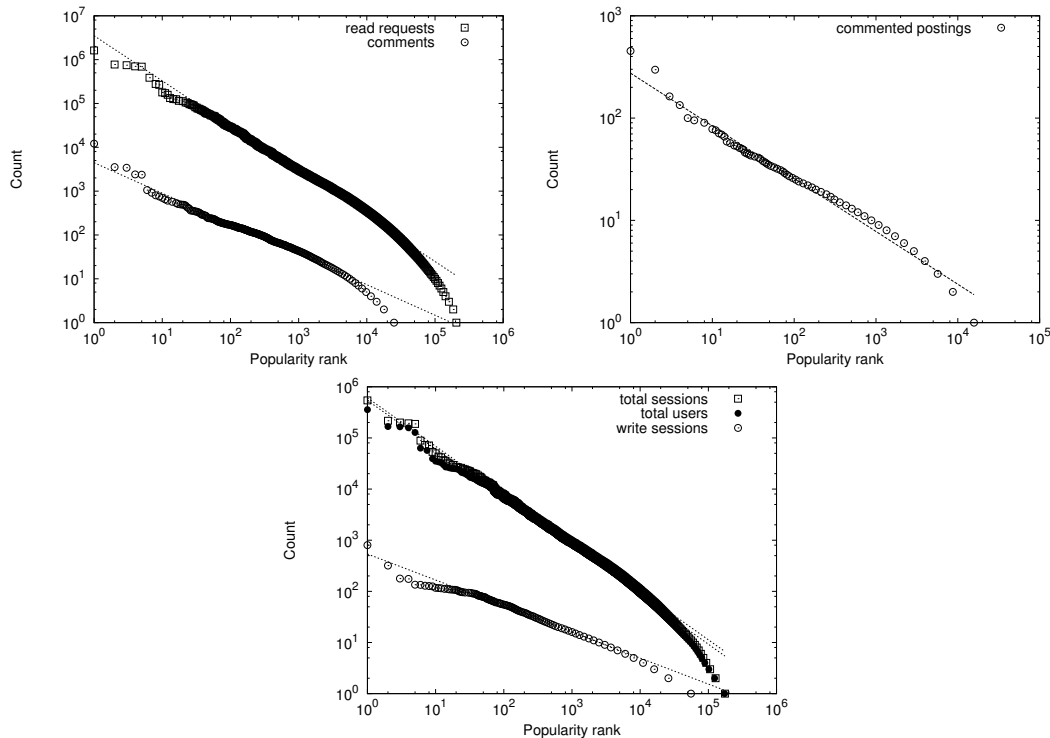


Fig. 2. Popularity profile of blogs: Frequency of read/write requests (top left), frequency of commented postings (top right); and frequency of total sessions, write sessions, and users (bottom) versus rank of blog. All profiles exhibit a power-law relationships.

4.2 Impact of Blogger Activity on Blog Popularity

Since various blogs elicit various degrees of interactions, it is natural to ask whether such interactions are correlated with the bloggers level of activity: *Does a high level of administrative activities for a blog imply a higher intensity of requests by visitors?* Figure 3 shows a scatter plot in which each blog is represented by a point showing the total number of sessions and the level of administrative activities for the blog. Figure 3 shows that the correlation between the level of administrative activities and overall number of sessions accessing a blog is quite weak (if any).

4.3 Blogs as Catalysts of User Interactions

As we alluded earlier, a major differentiating aspect of blogspace when compared to traditional web content is that in accessing blogosphere objects, users are in fact engaging in an exchange of postings and comments, which can be thought of as a dialogue between the various players – between the blogger and his/her readership as well as between members of the community catalyzed

by a given blog or set of blogs. In order to characterize the attributes of this dialogue, we propose the simple *dialogue structure* shown in Figure 4. In particular, such a dialogue can be seen as a sequence of postings by the blogger, read sessions and comments by visitors. Using these key blogger and visitor actions, we can define and characterize a number of attributes that allow us to quantify the levels of interaction induced by a given blog.

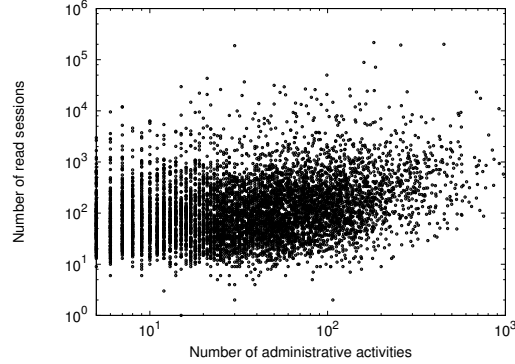


Fig. 3. Total sessions versus number of admin. requests.

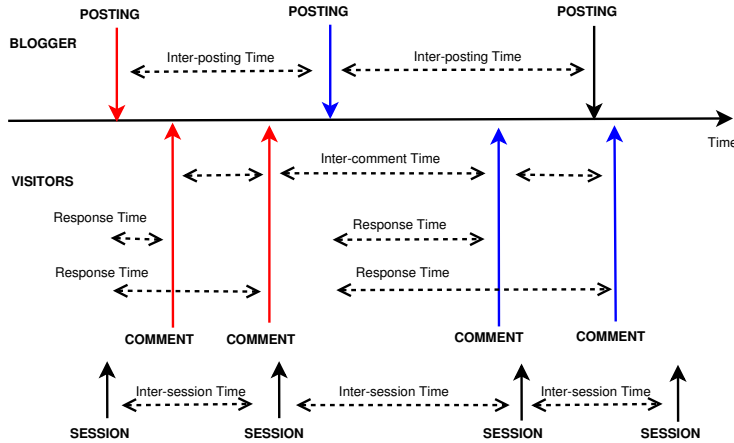


Fig. 4. The dialogue structure induced by a given blog is defined by the actions of the blogger and visitors through the interleaving of postings with sessions and comments.

One set of attributes that could be used to characterize the level of user engagement to a given blog are the interarrival times. In Figure 4 we can observe interarrival times of user sessions, postings, and comments. We refer to these by the *inter-session*, *inter-posting* and *inter-comment* times, respectively. Figure 5 show the CCDF of the marginal distributions of these times. Since read requests represent about 99% of visitors operations, is also interesting to study their interarrival times. The huge amount of *inter-read times* lead us to study this variable in different load periods, i.e., medium and heavy load. See section 7 for more details about our findings.

In addition to these interarrival times, another attribute of user interactions (enabled through a given blog) is the speed with which blogger postings elicit

feedback (*i.e.*, comments) from users. Figure 6 shows the distribution of the *response time*, which is defined as the time between a posting by a blogger and the various comments posted by visitors, as illustrated in Figure 4. Two key observations from that figure is that most (90%) of comments were received within one day of a new posting, and that hardly any comments were sent beyond one week of a posting.

For each attribute characterized in Figures 5 and 6, we show two distributions. The first is the aggregated distribution across all blogs, whereas the second is for the most popular blog in our blogosphere. In addition to the empirically observed distributions, Figures 5 and 6 also show the distributions that best fitted actual data, which were selected from the set of distributions commonly applied in the characterization of web workloads and presented in Table 2. Table 3 shows the best fitted distribution for each interaction attribute analyzed.

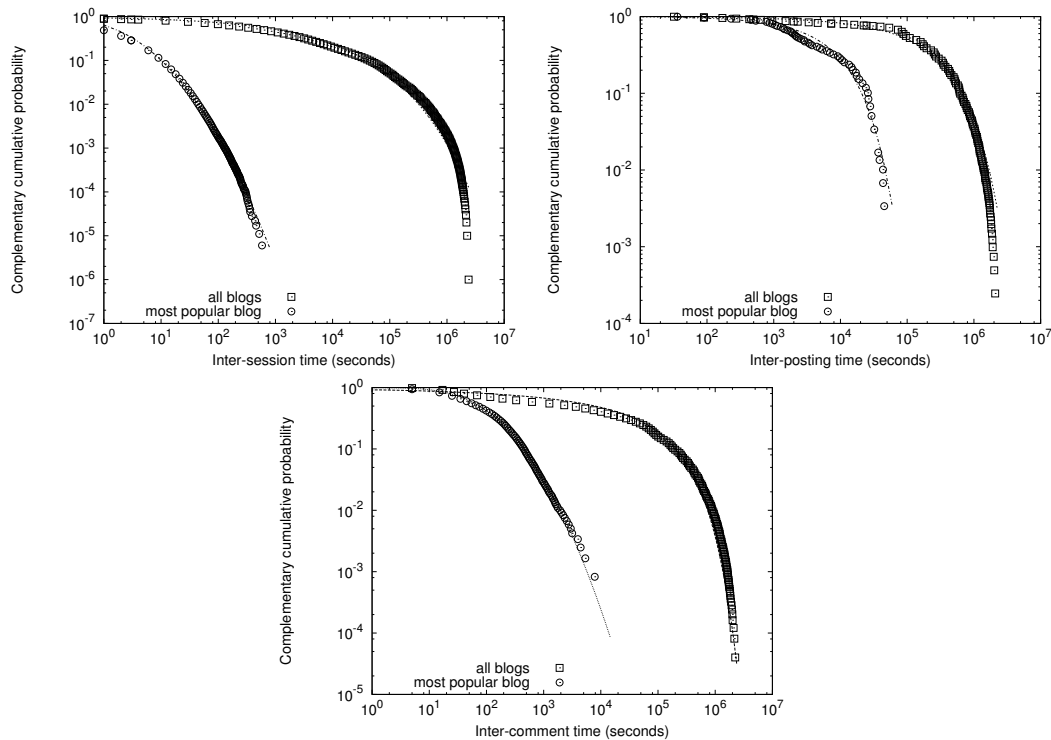


Fig. 5. Distribution of inter-session time (top left), inter-posting time (top right) and inter-comment time (bottom).

4.4 Classification of Blogs Based on Interaction Type

The previous discussion pointed out that accesses in a blogosphere could be seen as constituting dialogues (or sets of interactions) between blogosphere users – dialogues that are catalyzed by the blogs themselves. A natural question then is whether there is a difference in the type of interactions induced by the various blogs.

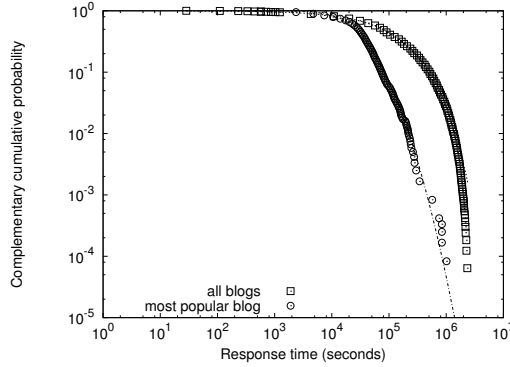


Fig. 6. Response time distribution.

Table 2

Distributions used in the characterization process.

Model	Probability Density Functions	Parameters
Pareto	$\alpha \frac{k^\alpha}{1-k^\alpha} x^{-\alpha-1}, k < x < 1$	α, k
Lognormal	$(1/\sigma x \sqrt{2\pi}) e^{-(\log(x)-\mu)^2/2\sigma^2}$	μ, σ
Gamma	$(1/\beta^\alpha \Gamma(\alpha)) x^{\alpha-1} e^{-(x/\beta)}$	α, β
Weibull	$\alpha \beta^{-\alpha} x^{\alpha-1} e^{-(x/\beta)^\alpha} I_{(0,\infty)}(x)$	α, β

One way of characterizing the interactions over a given blog is to quantify the intensity with which comments are posted to a given blog – *e.g.*, using the comment-to-request ratio as an indication of readership engagement. A blog with a low comments-to-request ratio underscores a blog that more or less features a one-way communication (or interaction) – *i.e.*, a blog through which the blogger “speaks” to his/her readership – much in the same way a newspaper editorial reflects a one-way communication between the author and readers. A blog with a high comments-to-request ratio underscores a blog that features a multi-way communication (or interaction) – *i.e.*, a blog through which the blogger as well as his/her readership are engaged in a multiway communication.

Figure 7 (top left) shows a scatter plot in which each point represents a blog. The coordinates of the blog reflect the number of sessions for that blog (on the x axis) and the number of write sessions for that blog, *i.e.*, number of sessions in which a comment was submitted to the blog. The scatter plot shows that a correlation exists (as one would expect) between the total number of sessions accessing a blog and the number of sessions posting a comment to the blog. However, the scatter plot also shows significant differences among blogs accessed by similar number of sessions. For instance, among blogs accessed by around 10,000 sessions, there is a blog which was accessed by only 2 write sessions (*i.e.*, only two user sessions resulted in comments being posted to the blog), whereas another was accessed by over 1,000 write sessions.

Table 3

Distributions and associated parameters best fitted to the various interaction attributes observed empirically.

Interaction	All blogs	Most popular blog
Attribute	Distr. and Parameters	Distr. and Parameters
Response Time	Weibull $\alpha = 0.64892$ $\beta = 0.00047$	Weibull $\alpha = 1.04838$ $\beta = 0.00002$
Inter-Session Time	Weibull $\alpha = 0.33081$ $\beta = 0.06963$	Lognormal $\mu = 0.516349$ $\sigma = 1.39864$
Inter-Posting Time	Gamma $\alpha = 0.46289$ $\beta = 528.04700$	Gamma $\alpha = 0.64255$ $\beta = 12.62400$
Inter-Comment Time	Gamma $\alpha = 0.20846$ $\beta = 328.57200$	Lognormal $\mu = 4.31054$ $\sigma = 1.40456$

Interestingly, the results in Figure 7 (top right) suggest that there is an inverse relationship between the likelihood of a blog being the object of posted comments by users (y axis) and the general popularity of the blog (x axis). Blogs on the right-hand-side of the plot in Figure 7 (top right) are those involving a large number of sessions, almost none of which are write sessions – they underscore a one-way, one-to-many “broadcast-like” communication from a very small number of writers to a large number of readers. This is much like the readership of a newspaper. On the other end of the scale, the blogs on the left-hand-side of the plot in Figure 7 (top right) are those involving a large number of write sessions – they underscore blogs that, while not too popular by virtue of total number of sessions accessing them, elicit comments from a large fraction of the visitors accessing them. This type of access is akin to that of a register (or guest-book, petition, *etc.*), for which the communication is many-to-one, and the purpose of access is to record a comment (or support a petition, *etc.*) Finally, blogs in the middle of the range in Figure 7 (top right) are those involving a fairly sizeable number of sessions, of which a non-trivial fraction of sessions are write sessions – they underscore popular blogs that elicit sizeable contributions from visitors. This type of access is akin to the exchanges in a parlor or public forum, in which the communication (while steered and/or moderated by a host) underscores a many-to-many dialogue between participants.

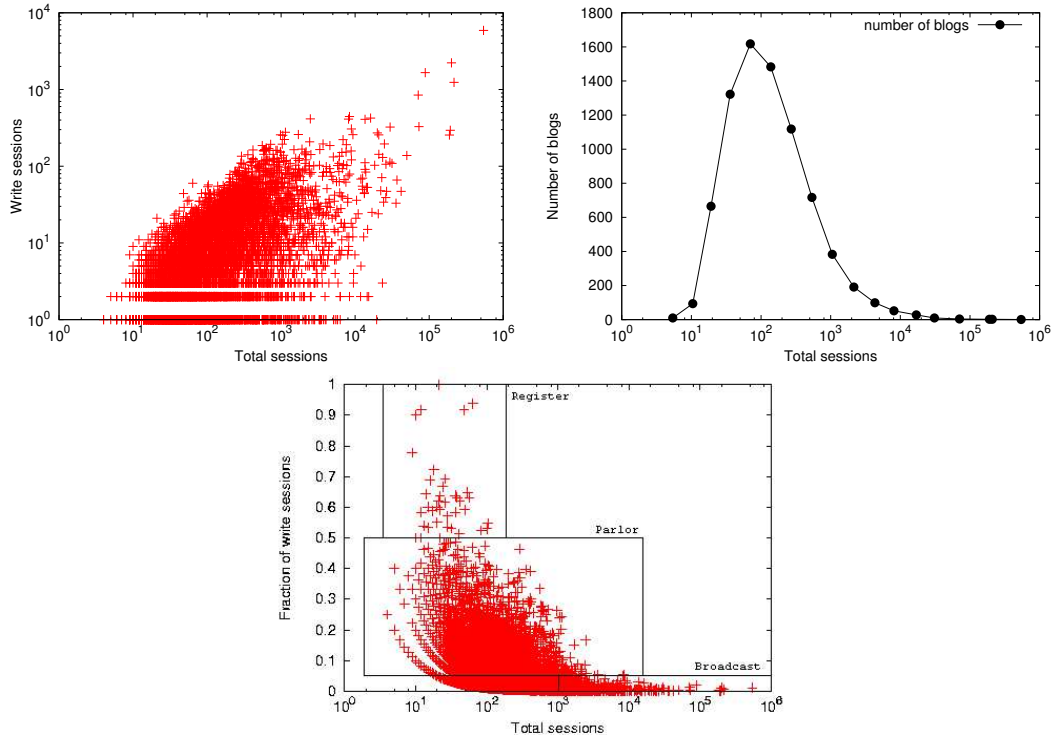


Fig. 7. Interactions induced by a blog as reflected by total number of sessions and the fraction of write sessions: Scatter plot showing correlation between total number of sessions and total number of write sessions (top left). Number of blogs and average ratio of write sessions for blogs with similar number of total sessions (top right). Scatter plot showing for each blog the ratio of write sessions versus the total number of sessions; blogs are classified as broadcast, parlor, or register blogs accordingly (bottom).

From the previous observations, we classified all blogs in our blogosphere into four categories based on their popularity and the ratio of write sessions they feature³. As illustrated in Figure 7 (bottom), *Broadcast-type blogs* are those accessed by more than 1,000 sessions, of which 5% or less of the sessions were write sessions. *Parlor-type blogs* are those for which more than 5% and less than 50% of all sessions were write sessions. *Register-type blogs* are those for which write sessions exceeded read-only sessions. Table 4 presents the resulting breakdown as observed in our blogosphere. The specific thresholds we used in our classification (namely 1,000 as a measure of intensity of access, and 5% and 50% as thresholds for the fraction of sessions featuring comments) were picked based on what we perceived as natural “clusters” of blogs in our blogosphere. Naturally, these thresholds and the resulting breakdowns would be different for other blogospheres, but the basic observation (and methodology) would hold.

³ Blogs with less than 50 sessions were excluded since there is not enough observations to support their classification.

Table 4

Breakdown of blogs/sessions by interaction type.

Blog Type	Percentage of		
	all blogs	all sessions	write sessions
Broadcast	7%	74%	25%
Parlor	55%	12%	63%
Register	1%	0%	1%
Unclassified	37%	14%	11%

5 Blogosphere Server View

In this section, we focus on the *server view* of the blogosphere workload, the lowest level of our hierarchy. The server view is the aggregation of accesses across all users and objects.

5.1 Marginal Distribution of Transfer Size

Figure 8 shows the Complementary Cumulative Distribution Function (CCDF) of the sizes of transfers from the blogosphere server. A Pareto distribution with parameter $\alpha \approx 1$ fits the transfer size distribution (see 2). This result is consistent with what has been observed for traditional web traffic [13,5].

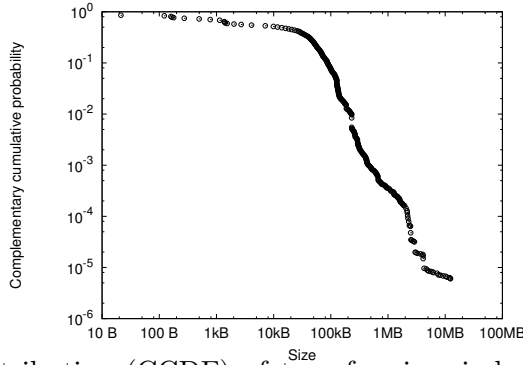


Fig. 8. Marginal distribution (CCDF) of transfer sizes is heavy-tailed, namely a Pareto with parameter $\alpha \approx 1$.

5.2 Diurnal Patterns

Looking at the server traffic over time, we observe that the intensity of the traffic induced by accesses in blogosphere follows a very strong diurnal pattern, with distinct peaks and valleys. Figure 9 illustrate this by showing the traffic

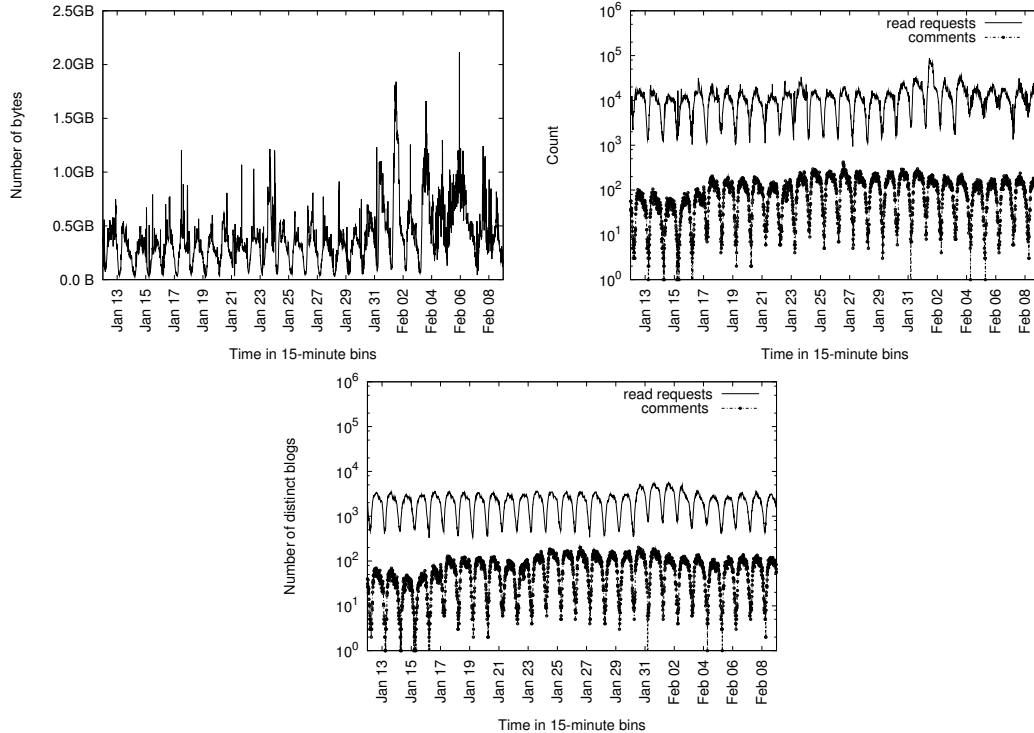


Fig. 9. Diurnal nature of blogosphere access patterns: Number of bytes transferred (top left), number of read/write requests sent by blogosphere visitors (top right) and number of distinct blogs accessed with read/write requests (bottom).

intensity at three distinct granular levels over time: measured in bytes, in number of requests and comments, and in number of unique blogs accessed (all measured in 15-minute intervals). The diurnal nature of blogosphere accesses is not dissimilar to that for general web content, as noted in a number of studies (see [6] for example). What is different for blogosphere workloads seems to be the high variability in the intensity of the peaks observed over time. This variability (which we document next) could be explained by noting that it is a byproduct of the bursty level of interactions between members of the community (or the social network) defined by a given blog, or set of blogs.

5.3 Burstiness of Peak Diurnal Access Intensity

The peak level of a diurnal access pattern depends mostly on the overall popularity of the content and on the fact that such popularity is a function of time (day, time-of-day, *etc.*).

For traditional web content, the change in the overall popularity of a web page tends to be smooth, which in turn results in low variability in the peaks observed in the diurnal access patterns (and mostly due to differences in intensity for weekdays versus weekends (see [6] for example). Exceptions to this

rule are web pages focusing on news and updates, pages linked by some very popular news websites or when their addresses are advertised to a wide public in the media [14,15]. In these cases the number of requests received by the web pages could grow rapidly, overloading the server capacity. Such events are often referred to as flash crowds [16]. In particular, blogspace content focusing on news and updates represent a distinct class of blogs that are typically popular, but which do not underscore/ elicit much blogosphere interactions. For blogosphere content, the popularity of a blog over time is more a function of the content of the blog (blog postings and comments, references from other popular blogs, *etc.*) as opposed to its “universal” popularity. To illustrate the impact of blog content on the popularity of a blog, Figure 10 shows the diurnal patterns of access (both read requests and comments) observed for a given blog – namely the *most popular* blog in our blogosphere. The figure underscores that the high variability in the peaks of the diurnal patterns (up to an order of magnitude for read requests) is not periodic (not weekdays versus weekends), but rather arbitrarily bursty. As an instance of this burstiness (modulated by diurnal patterns), one can observe a clear set of high-popularity periods – *e.g.*, the set of peaks starting with the peak on February 1st (also observed around January 14th, 19th and 24th). We have analyzed the diurnal patterns of administrative activities reflected in the *admin-log* (analysis not included due to space limitations) and have concluded that the surges in peak diurnal access intensities by visitors do *not* coincide with an intensity of new postings by the blogger. This leads us to conclude that it is the subject-matter of the postings (and not the mere number of postings) and the ensuing comments on and links from other blogs to these postings that results in these bursts. The impact of blogger activity on the blog popularity was previously analyzed in Section 4.2.

6 User Behavior Model

This section describes the user behavior model we propose to represent the activities of visitors in a blogosphere. This model, which is used as the basis for the design of our workload generator, GBLOT, includes only activities related to reading and commenting on the various blogs belonging to the blogspace, thus excluding administrative activities as they represent only less than 3% of our workload (see Table 1).

As a first step, the typical behavior of visitors of a blogosphere can be described, in high level, as follows: a user starts a new session by issuing a request to (read) a blog page. She may visit one or more entries in this blog as well as write comments thereon. Afterwards, the user may remain within the blogspace, visiting other blogs by following the various types of links that interconnect them. At some point, the user may either end the session or issue

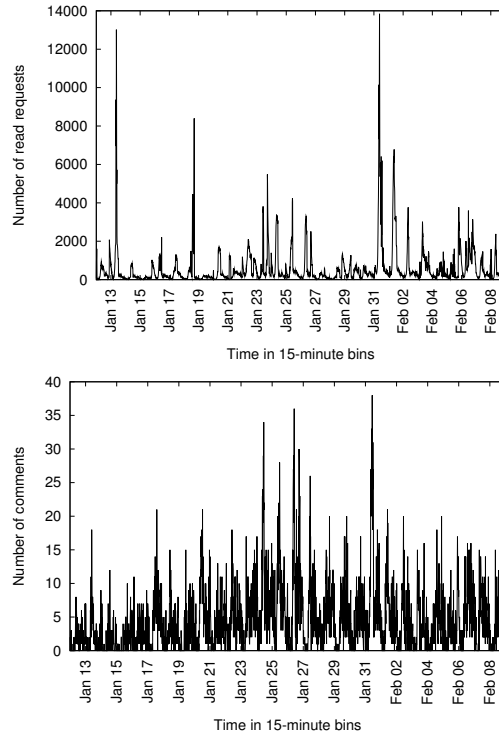


Fig. 10. Diurnal access patterns underscore high variability in popularity of a given blog over time: Number of read requests (top) and comments (bottom) to the most popular blog.

a request to a web page outside the blogspace, thus leaving it.

In order to model the behavior of a blogosphere visitor, including read and write request patterns to the various blogs visited within a session, we propose to use a *Visitor Behavior Model Graph* (VBMG), which is a state-transition graph proposed in [17]. In this graph, nodes represent possible states. A probability is assigned to each transition between two states. It is possible to characterize different types of users using VBMGs that differ in their transition probabilities. We define the following states a blogosphere visitor may be during a session:

Start Reading New Blog: A user visits this state either when she issues the first request to the blogosphere, thus entering it, or when she visits a new blog within the blogspace. A new blog could be visited after read or write requests to a previously visited blog.

Continue Reading Same Blog: A user visits this state, if, after reading a blog for a first time during a session, she decides to read other posts in that same blog. The user leaves this state when she either (1) reads a new blog, (2) decides to make a comment, or (3) navigates away by accessing a web page out of the blogspace.

Make Comments: A user visits this state when she writes comments on a blog post. This state is reached from any of the two states described above.

A user stays in this state as long as she continues to write comments to the same blog. The user leaves this state when she issues a read request to the same commented blog or to any other blog, or if she navigates away by accessing a web page out of the blogspace.

Exit: A session terminates when the user navigates away by accessing a web page out of the blogspace, or when the time since the last request exceeds a timeout value, which is assumed to be 30 minutes, in this paper.

Blog visitors can be grouped into different categories according to their visiting patterns. These categories are characterized by different VBMGs in terms of the state transition probabilities. For instance, the typical behavior of blurbers who tend to visit few blogs is represented by the VBMG shown in Figure 11. The behavior illustrated in the VBMG shown in Figure 12 is that of users who most frequently visit one (or very few) blog in the blogosphere (i.e., the probability of start reading a new blog is low), read and make comments in it and, eventually, leave the blogspace.

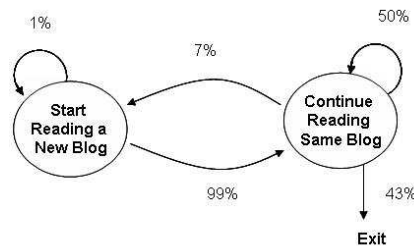


Fig. 11. VBMG of blurbers who tend to visit only a few blogs

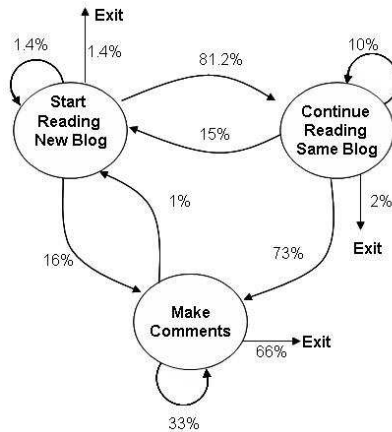


Fig. 12. VBMG of a typical commenter

6.1 Creating VBMGs from Server Logs

In order to create the VBMGs corresponding to users in our blogosphere, we use the following procedure, which is based on the algorithm described in [17].

The procedure receives as input a log file F with requests to the blog server, which we create by sorting and merging our write and read traces. As output, the procedure generates a session log S , with each record m (i.e., session) composed of a $n \times n$ matrix $C_m = [c_{i,j}]$ ($n = 4$) of transition counts between the four VBMG states for that session. The algorithm used to generate the *session log* S from the *request log* F is detailed next.

The request log file F is scanned sequentially. For each request, the algorithm determines if the request starts a new session or is part of an existing open session. A request starts a new session if there is no open session for the visitor who issued the request or if the time since the last request from the same visitor exceeded the timeout value (30 minutes). If the request starts a new session, the algorithm closes any open session from the same visitor, and sets the current state to be *Start Reading New Blog*. When the request is part of an open session, the next state is determined and the corresponding entry in the transition count matrix is updated. A session finishes and is closed when the user navigates away from the blogspace, at the end of the log file F , or if the session times out. Every time a session is closed a new entry is written to the session log.

After the session log S is built, the VBMGs corresponding to the sessions in the log are clustered into a number of visitor profiles, each corresponding to users who typically behave with a similar visiting pattern.

To that end, we employ the *k-means* clustering algorithm [18]. K-means clustering is an algorithm to group a given data set through a certain number of clusters (assume k clusters) fixed a priori. The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. The definition of distance presented in [17] is based on the transition count matrix. Given two points C_x and C_y in the session log, where each C_i is a transition count matrix, the Euclidean distance $d_{x,y}$ between them is defined as:

$$d_{x,y} = \sqrt{\sum_{i=1}^n \sum_{j=1}^n (C_x[i,j] - C_y[i,j])^2} \quad (1)$$

When the clustering algorithm includes a new point into a centroid, the centroid must be updated. If a point C_m is to be added to centroid k represented by point C , then the elements of the new centroid (C') are computed as

$$c'[i,j] = \frac{s(k) * c[i,j] + c_m[i,j]}{s(k) + 1} \quad (2)$$

where $s(k)$ is the number of points represented by centroid k .

Once all the clusters have been obtained, the procedure to get VBMGs derive the matrix P of transition probabilities associated with each cluster, as it is shown in equation 3.

$$p[i, j] = \frac{c[i, j]}{\sum_{k=1}^n c[i, k]} \quad (3)$$

The session arrival rate λ_k^s of members represented by the VBMG of cluster k is given by $\lambda_k^s = s(k)/T$, where T is the time interval during which the request log was obtained.

As proposed in [17], we determine the number of clusters based on the β_{cv} and β_{var} metrics. The β_{cv} metric is the ratio between the coefficient of variation computed for the intra-cluster and inter-cluster distances. Similarly, the β_{var} metric is the ratio between the intra and inter-cluster variances. The number of clusters is selected so as to reach the smallest values of β_{cv} and β_{var} .

6.2 Profiles of Visitors of our Blogosphere

Using the approach described in the previous section we uncovered six classes of visitors in the logs we analyze. The analysis was done in three days of medium and heavy traffic intensity. Since we made the analysis for periods of stability we opted for looking at each day separately. Visitors falling into classes 1, 2, 3 and 6 are mostly blurkers, as the number of write sessions from those users is less than 2%. Figure 11 shows the VBMG representing the profile of a typical **class-1** blurker. Classes 4 and 5 are composed by commenters, who have between 20% and 30% of write sessions. The VBMG corresponding to **class-4** visitors is shown in Figure 12.

Three aspects differ the typical behavior of blurker visitors in classes 1, 2, 3 and 6, namely, the exit probability, the probability of reading new blogs, and the probability of continuing reading the same blog. **Class-1** visitors tend to remain in the same blog throughout their visit, *i.e.*, the probability of reading new blogs is low (less than 10%), and have a high exit probability (43%). **Class-6** members have a similar blog visiting pattern. However, they have a much lower exit probability (around 1%), meaning that their sessions within the blogosphere tend to include a larger number of read requests. In contrast, **class-2** and **class-3** blurkers read new blogs with higher probabilities. In particular, transition probabilities from state *Continue Reading Same Blog* to *Start Reading a New Blog* are 76% and 32% for classes 2 and 3, respectively. Moreover, the exit probability of **class-3** visitors is higher than that of **class-2** (about 50% vs 8%, respectively).

In conclusion, **class-2** and **class-3** blurkers tend to read many different blogs when compared to those users of **class-1** and **class-6** categories. Moreover, **class-2** and **class-6** visitors tend to send a larger number of read requests (they have a lower exit probabilities) in comparison with those falling into classes 1 and 3.

Visitors falling into classes 4 and 5 typically write comments during their sessions. However, **class-5** visitors tend to be even more interactive and to remain in the blogosphere longer. Whereas **class-4** commenters only write a few comments in one blog (the probability of reading new blogs is less than 20%) and exit with a probability of 66%, **class-5** visitors have higher probabilities of making more than one comment in the same blog (76% vs. 33% of **class-4** members), and of reading many different blogs during the same visit (about 46%).

Based on the characteristics of different types of blogs, we find that most members of classes 1,2, 3 and 6 could be viewed as visitors of Broadcast-type blogs. In comparison, members of classes 4 and 5 are more interactive, and thus, should typically visit Parlor-type and Register-type blogs.

7 GBLOT: A Generator of Blogosphere Workloads

The generation of *realistic* synthetic blogosphere workload and traffic is of great importance for capacity planning of blog content hosting sites, and for traffic engineering purposes. Several tools have been developed to generate representative HTTP workloads ([19–22]). However, HTTP request generators are not suitable for generating synthetic blogosphere traffic as they do not capture key aspects of the interaction of users and blogspace content, which is different from how user accesses to traditional Web pages.

Therefore, we designed and implemented a generator of synthetic blogosphere workloads called (GBLOT)⁴, which makes use of the various statistical characteristics and models presented in this paper. The workload generation model adopted in GBLOT is based on our characterization of user sessions presented in Section 6, whereby the characteristics of representative classes of visitors are used to synthesize the blogosphere workload. GBLOT was designed under the assumption of stationarity – *i.e.*, the workload parameters remain stationary for the period of time spanned by the synthetically generated workload. In particular, dynamic blog properties that are not considered in GBLOT include: changes (growth) in the size of various blogs due to the introduction of new postings by the blogger, blog popularity temporal variation, and workload

⁴ The tool can be downloaded from <http://www ldc.usb.ve/~gblot>.

variability due to diurnal access patterns. We also note that although the tool was conceived primarily to generate visitor workloads, its design allows for the generation of administrative workloads as well.

7.1 GBLOT: Inputs and Outputs

GBLOT receives as input a description of the synthetic workload to be generated. The workload description includes the *number of blogs* that comprise the target blogosphere, the *duration* (in seconds) of the synthetic workload, the *number of classes of blogosphere visitors*, and a *class description file*, containing the characteristics of the behavioral model of that class, namely, the distribution parameters that define the inter-session times, and the transition probabilities among the different states of that visitor class (i.e., the VBMG characteristic of the class).

Once completed, GBLOT's synthetic workload generation process produces two main outputs: a *workload summary file* and a *synthetic access log file*. The summary file provides an overview of the generated trace, with overall statistics of its key attributes, such as the number of sessions for each visitor class, the number of accesses to each blog, the number and percentage of read and comment requests, among others. The synthetic access log file contains records describing per-user operations (i.e., requests) within the blogosphere. Each such record contains a timestamp, a visitor identifier (including her class), a blog identifier and the type of the request (read or comment). Notice that requests from various visitors are merged in the synthetic access log, which allows it to be used directly for trace-driven evaluations of blog server performance.

7.2 GBLOT: Design

GBLOT's software architecture follows the architecture of the Surge HTTP workload generator [19]. It consists of three main components: a set of programs to generate distributional data, a master program and a multiprocess-multithreaded program, which, jointly, generate the concurrent visitor sessions that will make up the target synthetic workload.

The master process creates a *g-process* for each visitor class. Each *g-process* implements a visitor class according to the corresponding *visitor class description file* (i.e., its VBMG). To do so, it dynamically creates threads to *simulate* concurrent sessions submitting requests to the blog server. The threads are created following the session inter-arrival time distribution. Each thread is responsible for generating the requests that compose a single session. To that

end, each *g-process* requires as input parameters, in addition to the visitor class description file, the total number of unique blogs, distributional data describing blog accesses, and the target output file descriptor. The master process interrupts the generation of new sessions (and requests) once the specified total workload *duration* is reached.

GBLOT also includes a set of tools to generate distributional data, according to statistical models identified in Sections 5 and 4. GBLOT includes distributions of blog popularity (in terms of the total number of read and write requests) and transfer sizes. Blog requests that follow the identified distributions are generated and stored in a file before the execution of the master process. The master process feeds these data into a nominal pipe which provide input data to the threads simulating visitor sessions. Note that, by generating the distributional data separately from the sessions, we separate the visitor behavior model from the other workload aspects (e.g. popularity, file transfer sizes), thus making it easy to experiment with alternative statistical models for these aspects.

The number of threads to be created is limited only by the maximum number of threads per process allowed by the operating system. We use kernel threads to avoid blocking when the parent process sleeps to simulate the session inter-arrival times, thus providing a higher degree of concurrency. Moreover, individual kernel threads can take advantage of multiprocessors. GBLOT is implemented using the C language.

7.3 *GBLOT: Validation*

This section provides a validation of GBLOT by comparing some of the characteristics of synthetically generated workloads with those observed in the real workload. The characteristics considered in the comparison include not only workload attributes directly used in GBLOT's parametrization, such as blog popularity, but also other attributes that are indirectly derived, such as the inter-read times, inter-comment times, and inter-read times for the most popular blog (referred to as MPB). Inter-read (inter-comment) times refer to the time intervals between two consecutive read (write) requests within the same visitor session.

7.3.1 *Parameterization*

Recall that GBLOT is based on the assumption of stationary workloads. Thus, instead of extracting its input parameters from our entire blogosphere workload, which spans a 28-day period, we looked into the workloads observed in individual days. We analyzed a number of different days, and selected results

of two days that are representative of *heavy* and *medium* workloads. Table 5 shows the number of sessions as well as the numbers of read requests and comments in each day.

Table 5

Workload trace characteristics used in the validation.

	# Number of Sessions	Number of Read Requests	Number of Comments
Medium Load	202172	899509	9262
Heavy Load	592424	2944464	11801

The procedure described in Section 6 was applied to the workload traces of both days to obtain the parameters of the VBMGs representing the visitor classes. As mentioned in Section 6, our analysis of the entire blogosphere traces resulted in the identification of six classes of visitors. However, classes 5 and 6 were very rarely observed in the two days considered, accounting for only a tiny fraction of the sessions. Therefore, for the purpose of validating GBLOT’s synthetic traces, we focus only on classes 1 through 4, which accounted for at least 1% of the sessions in either of the two days.

In order to obtain the parameters of the distributions of session inter-arrival time for each class, we characterized the number of sessions from each class observed during hourly periods of roughly stable arrival rates in both days. We found that, for all classes, session arrival counts in each analyzed period can be well described by a Poisson process⁵ with parameter λ , the session arrival rate. Table 6 gives an overview of the classes analyzed, showing, for each class, the daily percentage of visitors, the average number of states visited in a session, referred to as the session length \bar{S} , and the session arrival rate λ .

Table 6

Summary of the visitor classes.

Classes	Medium Load			Heavy Load		
	% of Sessions	Session Length (\bar{S})	λ	% of Sessions	Session Length \bar{S}	λ
Class 1	73	5.12	2.14	73.94	4.49	3.67
Class 2	2.15	52.78	0.13	6.29	26.79	0.27
Class 3	21.77	8.58	0.63	16.99	3.03	1.06
Class 4	1.71	2.90	0.15	0.55	2.53	0.15

⁵ The goodness-of-fit was tested using as a tool the robust distribution plot, a graphical method for discrete data described in [23].

The number of unique blogs in the synthetic workloads was set to 3400, the average number of blogs accessed during the selected evening hour in both days. Moreover, since the popularity of objects in the blogspace follows a power law (as shown in Figure 13), the sequence of visited blogs in our target workloads was generated using a Zipf Distribution[24]. Similarly, the sequence of transfer sizes was generated using a Pareto distribution, consistently with our characterization results.

GBLOT was then used to generate two synthetic workloads, for medium and heavy load periods, both with durations equal to one hour. The characteristics of synthetically generated workloads were then compared against the corresponding one-hour *real* traces they were built upon. Table 7 shows the percentages of read and comments as well as the total number of sessions for synthetic and real workloads in each analyzed day. Figure 13 and Tables 8 through 11 show the parameters and other statistics of the best-fitted distribution models for each analyzed workload metric. Some of the empirical cumulative distribution functions are shown in Figures 14 and 15, to illustrate the agreement between synthetic and real workloads.

Whereas GBLOT produced fractions of read and comment requests similar to those observed in the real workloads, the numbers of sessions in the GBLOT synthetic trace are somewhat smaller than the ones observed in the real traces.⁶ Nevertheless, Figure 13 shows that the distribution of blog popularity in each synthetic workload agrees reasonably well with the distribution in the corresponding trace. The slopes of the two pairs of curves, presented on log-log scales, are very close and both are quite close to the skew parameter presented in Section 4 for the distribution of the number of read requests per blog.

Moreover, Tables 8, 9, 10, 11 show that the best-fitted distribution models for the other three workload attributes analyzed, namely, inter-read times, inter-comment times and inter-read times for the most popular blog (MPB) are the same for synthetic and real workloads, in both days. In addition, distribution parameters, minimum, mean and maximum as well as distribution quartiles are very similar in all cases. Finally, the empirical distributions of inter-read and inter-comment times shown in Figure 14 as well as the distribution of inter-read times for the most popular blog, shown in Figure 15 illustrate that GBLOT is able to capture reasonably well key workload aspects that are not directly represented in the session and workload models it relies on.

In summary, GBLOT implements a model of user behavior based on VBMGs, parameterized with distributional models presented in this paper. The results

⁶ This could be due to limitations related to the maximum allowable number of threads per process and the sharing of file access (in the nominal pipe), which could be delaying the termination of existing sessions and the creation of new ones.

Table 7

Summary of synthetic and corresponding real workloads (8pm-9pm in each day).

Day	% of Reads		% of Comment		# of Sessions	
	Real	Synthetic	Real	Synthetic	Real	Synthetic
Medium Load	98.89	98.40	1.11	1.60	9824	7674
Heavy Load	99.15	98.09	0.85	1.91	16833	12701

Table 8

Distributional models and associated parameters for different workload attributes (Medium Load Day).

Workload Attribute	Real Data	Synthetic Data
	Dist. and Parameters	Dist. and Parameters
Inter-Read Time	Lognormal $\mu = -2.53479$ $\sigma = 0.14092$	Lognormal $\mu = -2.37924$ $\sigma = 0.21540$
Inter-Comment Time	Gamma $\alpha = 1.22192$ $\beta = 6.33801$	Gamma $\alpha = 0.75396$ $\beta = 6.55213$
Inter-Read Time (MPB)	Gamma $\alpha = 1.06857$ $\beta = 2.50429$	Gamma $\alpha = 1.00748$ $\beta = 2.04545$

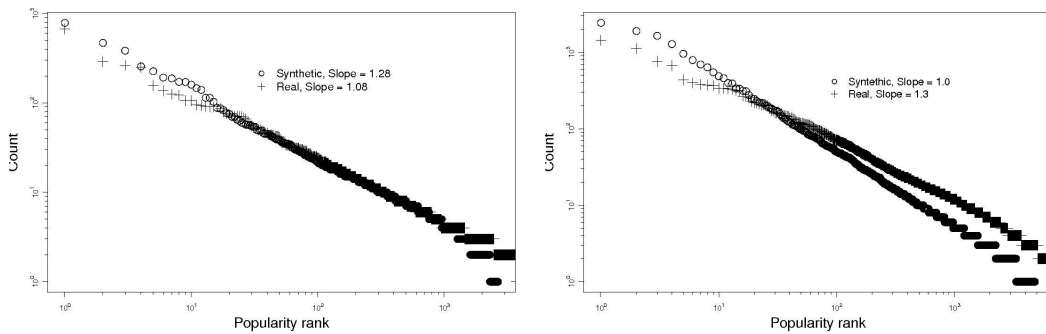


Fig. 13. Popularity profile in synthetic and real workloads: Number of read and write requests versus rank of blog. Medium Load Day (left), Heavy Load Day (right).

presented in this section show that GBLOT captures significant characteristics of workloads generated by visitors. In particular, it is able to generate read and comment requests with temporal characteristics that follow the same

Table 9
Summary of fitted distributions (Medium Load Day).

Interaction Attribute		Min.	1st Q.	Med.	Mean	3er. Q.	Max
Inter-Read Time	Real	0.054	0.073	0.080	0.081	0.089	0.12
	Synthetic	0.052	0.083	0.093	0.096	0.109	0.179
Inter-Comment Time	Real	0.004	2.864	6.182	8.364	11.490	54.760
	Synthetic	0.001	0.944	3.131	4.962	6.690	45.780
Inter-Read Time (MPB)	Real	0.002	0.577	1.475	2.097	2.895	15.260
	Synthetic	0.002	0.809	1.845	2.749	3.948	17.660

Table 10
Distribution models and associated parameters for different workload attributes (Heavy Load Day).

Workload Attribute	Real Data	Synthetic Data
	Dist. and Parameters	Dist. and Parameters
Inter-Read Time	Weibull $\alpha = 9.70175$ $\beta = 0.04681$	Weibull $\alpha = 3.55986$ $\beta = 0.11109$
Inter-Comment Time	Gamma $\alpha = 1.14587$ $\beta = 4.96327$	Gamma $\alpha = 1.00267$ $\beta = 4.09847$
Inter-Read Time (MPB)	Weibull $\alpha = 0.82184$ $\beta = 2.33241$	Weibull $\alpha = 0.96391$ $\beta = 1.61360$

distributional characteristics observed in real traces.

Table 11

Summary of fitted distributions (Heavy Load Day).

Interaction Attribute		Min.	1st Q.	Med.	Mean	3er. Q.	Max
Attribute							
Inter-Read Time	Real	0.094	0.308	0.378	0.396	0.489	0.778
	Synthetic	0.067	0.251	0.356	0.404	0.512	1.426
Inter-Comment Time	Real	0.005	1.726	4.188	5.598	7.908	29.530
	Synthetic	0.010	1.211	2.850	4.152	5.532	36.460
Inter-Read Time (MPB)	Real	0.0004	0.493	1.087	1.711	2.304	13.430
	Synthetic	$2.07e^{-5}$	0.482	1.438	2.435	3.230	2.195

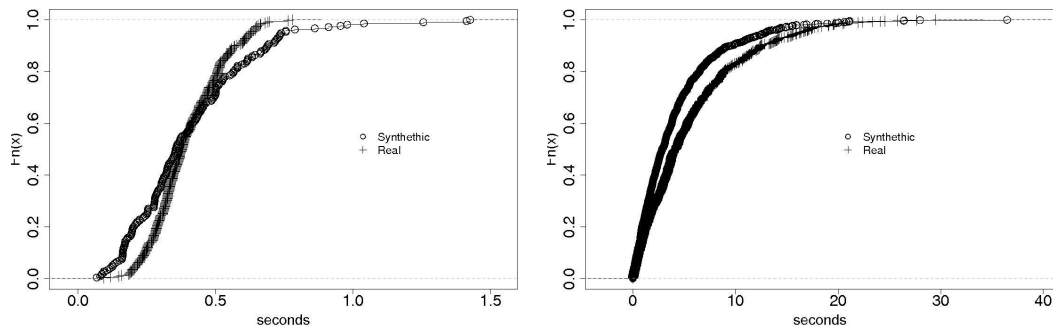


Fig. 14. Empirical cumulative distribution functions (Heavy Load Day): inter-read time (left), inter-comment time (right).

8 Related Work

Workload characterization is fundamental to the understanding and engineering of Internet systems. Many studies focused on the characterization of Internet traffic and web access workloads; examples of key early studies along these lines include [25,11,26–30]. Some of the important findings of these studies

include establishing the Zipf-like popularity of web objects, the heavy-tailed object and request size distributions, and the temporal and spatial reference locality in request streams. A discussion of the various characteristics of work-

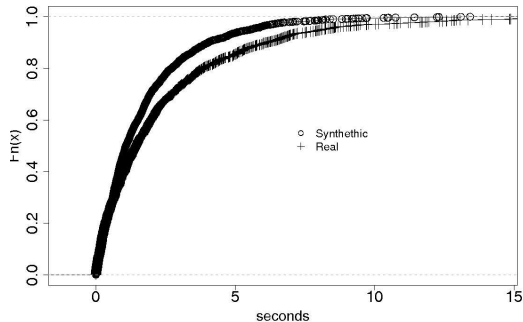


Fig. 15. Empirical cumulative distribution function (ECDF): inter-read time in the heav loaded day (most popular blog).

loads involving traditional web content (while relevant to some aspects of our work) is outside the scope of this paper. With respect to workload generators, [19–21] and [22] describe some representative tools for HTTP workloads. However, none of them capture key aspects of the dynamics of user interactions in the blogosphere, in particular commenting or posting. We are not aware of any previous generator of blog workloads. Thus, in the remainder of this section, we restrict our coverage of related work to studies that focused on modeling and characterization of blogspace workloads.

In our work, we used salient statistical features of blog access patterns to classify (or infer) the type of interactions that are facilitated by a blog among blogosphere actors. Along these lines, there has been a number of studies that used blog and blogger characteristics for inference purposes. For example, Kolari, Java, and Finin [31] examined the “Splogosphere” of spam blogs (splogs) used to host spam postings, whose purpose is to inundate blog search engines and blog search pinging services. They provided a comparison of the characteristics of authentic blogs with splogs, which could be used to differentiate them. Nakajima *et al* proposed a method for identifying bloggers who take on an important role in blog conversations (or threads) within a blogosphere [32]. Herring *et all* showed that the interconnections between blogs can be used to characterize relationships between blogs, and to infer clusters of conversations and communities [33].

The results presented in this paper stand in some contrast to those observed in [34] based on a microscopic analysis of a much smaller set of 724 blogs from different blog services. In that work, Mishne and Glance noted the correlation between popularity (measured statically using number of incoming links or dynamically using number of page views) and the number of comments posted to a blog. They also noted the existence of *outliers* – popular blogs that elicit noticeably small number of comments – which they attributed to blogger moderation or censorship. Since our traces allow us to measure the number of submitted comments (as opposed to just those approved by the

blogger), we are able to conjecture that the existence of highly popular blogs with relatively small number of comments is *not* a byproduct of blogger moderation, but is in fact a characteristic of a special class of blogs that act as conduits for one-to-many, “broadcast-type” interactions. Indeed, our results also show that the correlation between popularity and the number of posted comments does not hold even for less popular blogs, some of which may elicit a relatively large number of comments, acting as conduits for many-to-one, “register-type” interactions.

In [4], Cohen and Krishnamurthy noted that blogs provide a multi-way communication paradigm that regular web pages do not. Their analysis of a popular blogspace server showed that the rate of change of blogs is quite different from traditional web pages and that the nature and count of links between blogs and other web pages are quite distinct. They provided simple heuristics for inferring whether a web page is a blog or not, and they showed that tracking a seed collection of blogs could be used to identify emerging interests or on-line dialogues [35].

In [8], Kumar *et al* considered the temporal evolution of blogspace as an instance of hyperlinked corpora, noting the bursty nature of its evolution patterns, and highlighting the possibility of automatic community identification and burst extraction. In [7], Adar *et al* argued that such bursts could be traced to two lower levels of interactions, at the blogger level and at interest group level. In [2], Gruhl *et al* investigated the dynamics of information propagation through this hyperlinked corpus by identifying and tracking discussion topics using a “chatter and spikes” model, and then using biologically-inspired infectious disease propagation models to follow the diffusion of such discussions through blogspace.

9 Conclusion

In this paper we used an extensive set of traces to hierarchically characterize the access patterns in blogspace. Three levels were analyzed: user, object and server level. In addition to providing statistical models for various characteristics (popularity profiles, interarrival times, *etc.*), our study has unveiled a number of interesting findings, some of which are different from those widely accepted for traditional web content. These findings include our conjecture that access to objects in blogspace underscores an interaction between authors and a readership community, which can be classified based on blog popularity and read/write access characteristics as broadcast, parlor, or register interactions, and our conjecture that unlike traditional web pages, blogosphere access patterns are much more dependent on the social networks that they catalyze.

The workload models obtained by the characterization process were used to design and implement GBLOT, a synthetic workload generator – the only blogosphere workload generator of which we are aware. It simulates the behavior of several visitors within the blogosphere. GBLOT’s user session model is based on VBMGs (Visitor Behavior Model Graphs). The access patterns and transfer sizes in these sessions follow parametrized distributional models, based on characteristics of real traces. The synthetic workloads generated by GBLOT were shown to be quite similar to real workloads.

References

- [1] A. Ntoulas, J. Cho, C. Olston, What’s new on the web? the evolution of the web from a search engine perspective, in: Proc. of 13th International World Wide Web Conference, ACM Press, 2004, pp. 1–12.
- [2] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information diffusion through blogspace, in: Proc. of 13th International World Wide Web Conference, ACM Press, 2004, pp. 491–501.
- [3] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, Structure and evolution of blogspace, *Communications of ACM* 47 (12) (2004) 35–39.
- [4] E. Cohen, B. Krishnamurthy, A short walk in the blogistan, *Computer Networks* 50 (5) (2006) 615–630.
- [5] M. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: Evidence and possible causes, *IEEE/ACM Trans. on Networking* 5(6).
- [6] E. Velloso, V. Almeida, W. Meira, A. Bestavros, S. Jin, A hierarchical characterization of a live streaming media workload, *IEEE/ACM Transactions on Networking* 14 (1) (2006) 133–146.
- [7] E. Adar, L. Zhang, L. Adamic, R. Lukose, Implicit structure and the dynamics of blogspace, in: Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, 2004.
- [8] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, On the bursty evolution of blogspace, in: Proc. of 12th International World Wide Web Conference, ACM Press, 2003, pp. 568–576.
- [9] Giant blogging terms glossary: Need a blog dictionary?, <http://www.quickonlinetips.com/archives/2006/06/the-giant-blogging-terms-glossary> (2006).
- [10] D. Menasce, V. A. Almeida, R. Riedi, F. Ribeiro, R. Fonseca, W. Meira, In search of invariants for e-business workloads, in: Proceedings of the 2000 ACM Conference in E-commerce, 2000, pp. 56–65.

- [11] C. Cunha, A. Bestavros, M. Crovella, Characteristics of WWW client-based traces, Tech. Rep. BU-CS-95-010, Computer Science Department, Boston University (April 1995).
- [12] R. Jain, The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling, Wiley-Interscience, New York, NY, 1991.
- [13] A. Williams, M. Arlitt, C. Williamson, K. Barker, Web workload characterization: Ten years later, in: Web Content Delivery, Springer, 2005.
- [14] First statement of the food agency in belgium (afsca), <http://www.influenza.be/fr/persberichten/2005-10-27-afsca-obligation-de-confinement.doc> (2005).
- [15] Second statement of the food agency in belgium (afsca), <http://www.influenza.be/fr/persberichten/2005-10-28-afsca-obligation-de-confinement.doc> (2005).
- [16] U. Herman-Izycka, Flash Crowd Prediction, Master’s thesis, Vrije Universiteit, Amsterdam, The Netherlands (2006).
- [17] D. Menascé, V. Almeida, R. Fonseca, M. Mendes, A methodology for workload characterization of e-commerce sites, in: Proc. of 1st ACM conference on Electronic Commerce, ACM Press, 1999, pp. 119–128.
- [18] J. Hartigan, M. A. Wong, A K-Means Clustering Algorithm, Applied Statistics 28 (1) (1979) 100–108.
- [19] P. Badford, M. Crovella, Generating representative web workloads for network and server performance evaluation, in: Proc. of ACM SIGMETRICS’98, ACM Press, 1998, pp. 151–160.
- [20] G. Banga, P. Druschel, Measuring the capacity of a web server, in: Proc. of USENIX symposium of Internet Technologies and Systems, Monterrey, CA, 1997, pp. 61–71.
- [21] D. Mosberger, T. Jin, httpperf: A tool for measuring web server performance, in: First Workshop on Internet Server Performance, ACM, 1998, pp. 59–67. URL citeseer.nj.nec.com/mosberger98httpperf.html
- [22] A. Rousskov, D. Wessels, High Performance Benchmarking with Web Polygraph, Software–PRACTICE AND EXPERIENCE 1 (2003) 1–10.
- [23] M. Friendly, Visualizing Categorical Data, SAS Publishing, 2001.
- [24] J. Gray, P. Sundaresan, S. Englert, K. Baclawsky, P. Weinberger, Quickly generating billion-record synthetic databases, ACM SIGMOD Record 23 (2) (1994) 243–252.
- [25] V. Paxson, Wide-area traffic: The failure of Poisson modeling, in: Proceedings of SIGCOMM, 1994, pp. 257–268.

- [26] V. Almeida, A. Bestavros, M. Crovella, A. de Oliveira, Characterizing reference locality in the WWW, in: Proceedings of PDIS, 1996, pp. 92–107.
- [27] M. Arlitt, C. Williamson, Web server workload characteristics: The search for invariants, *IEEE/ACM Trans. on Networking* 5(5).
- [28] S. D. Gribble, E. A. Brewer, System design issues for Internet middleware services: Deductions from a large client trace, in: Proceedings of USENIX Symposium of Internet Technologies and Systems, 1997, pp. 19–19.
- [29] P. Barford, A. Bestavros, A. Bradley, M. Crovella, Changes in Web client access patterns: Characteristics and caching implications, *World Wide Web* 2(1) (1999) 15–28.
- [30] L. Breslau, P. Cao, L. Fan, G. Phillips, S. Shenker, Web caching and Zipf-like distributions: Evidence and implications, in: Proceedings of INFOCOM, 1999, pp. 126–134.
- [31] P. Kolari, A. Java, T. Finin, Characterizing the splogosphere, in: Workshop on the Weblogging Ecosystem, 15th International World Wide Web Conference, 2006.
- [32] S. Nakajima, J. Tatemura, Y. Hino, Y. Hara, K. Tanaka, Discovering important bloggers based on analyzing blog threads, in: Workshop on the Weblogging Ecosystem, 14th International World Wide Web Conference, 2005.
- [33] S. Herring, I. Kouper, J. Paolillo, L. Scheidt, M. Tyworth, P. Welsch, E. Wright, N. Yu, Conversations in the blogosphere: An analysis from the bottom up, in: Proc. of 38th Hawaii International Conference on System Sciences (HICSS'05), 2005, p. p107b.
- [34] G. Mishne, N. Glance, Leave a replay: An analysis of weblog comments, in: Workshop on the Weblogging Ecosystem, 15th International World Wide Web Conference, 2006.
- [35] J. Eckmann, E. Moses, D. Sergi, Entropy of dialogues creates coherent structures in e-mail traffic, *Proc. of the National Academy of Sciences* 101 (40) (2004) 14333–14337.