Traffic Characteristics and Communication Patterns in Blogosphere

Fernando Duarte[†] fernando@dcc.ufmg.br
Bernardo Mattos[†] bemattos@dcc.ufmg.br
Azer Bestavros[‡] best@cs.bu.edu
Virgilio Almeida[†] virgilio@dcc.ufmg.br
Jussara Almeida[†] jussara@dcc.ufmg.br

[†]Computer Science Department, Federal University of Minas Gerais, Brazil
[‡]Computer Science Department, Boston University, Massachusetts, USA

Abstract

We present a thorough characterization of the access patterns in blogspace – a fast-growing constituent of the content available through the Internet – 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 administrative requests spanning a 28-day period is done from three different blogosphere perspectives. The *server view* characterizes the aggregate access patterns of all users to all blogs; the *user view* characterizes how individual users interact with blogosphere objects (blogs); the *object view* characterizes how individual blogs are accessed. Our findings support two important conclusions. First, we show that the nature of interactions between users and objects is fundamentally different in blogspace than that observed in traditional web content. Access to objects in blogspace could be conceived as part of an interaction between an author and its readership. As we show in our work, such interactions range from one-to-many *"broadcast-type"* and many-to-one *"registration-type"* communication between an author and its readers, to multi-way, iterative *"parlor-type"* dialogues among members of an interest group. This more-interactive nature of the blogosphere leads to interesting traffic and communication patterns, which are different from those observed in traditional web content. Second, we identify and characterize novel features of the blogosphere workload, and we investigate the similarities and differences between typical web server and blogosphere server workloads.

Keywords

Blogosphere, Workload Characterization, Social Interaction, Communication Patterns

1. Introduction

A distinct and rapidly growing component of the web of 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 audience (*e.g.*, personal diaries, instructor notes for a 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 and may elicit further comments by other blog readers, and may trigger the addition of new entries in the same blog or in other blogs, *etc.* As such, blogs are in fact snapshots of an interactive exchange between the 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 [16], 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.

1.1 Motivation: The last few years have witnessed a significant growth in the size of the blogspace. In 2002, Newsweek estimated the number of weblogs to be half a million, attributing the "explosion" of blogspace size to the Blogger.com service (now part of Google) [11]. As of November 2006, the blogspace reached a phenomenal size of over 60 million blogs [20] – a whopping 120-fold growth over four years.

Given the prominence and continued growth of blogspace, it is natural to ask whether its characteristics are similar to those of more traditional components of the web. Indeed, over the last few years, there have been a number of studies that explored the various aspects of the blogspace. For example, the works in [9, 13, 5] 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 accesses 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 has uncovered important properties that were crucial in explaining observed traffic characteristics [6], and were instrumental in building workload models and in developing synthetic traffic generation tools [18]. 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 [1] or the evolution of its topological structure [13]).

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* [13]. We use the term *blogosphere* to refer to blogspace and the community of users (and underlying social networks) accessing it [10]. We use the term *blogger* to refer to the owner of a blog, an individual who keeps and updates 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. We differentiate between a *write session* and a *read session* (or simply *session*) based on whether, in addition to accessing the blog, the visitor submitted a comment or not.

1.3 Goals and contributions: Based on an extensive workload from a major Internet Service Provider (ISP) in Brazil – a workload that consists of over 32 million requests to over 210,000 blogs that resulted in almost one TeraByte of transfered content over a 4-week period – we provide a statistical analysis of how users read blogs and send 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 than those observed in access patterns of traditional web content. In that respect, we identify and characterize novel features of the blogosphere workload, and we investigate the similarities and differences between typical web server workloads and blogosphere server workloads.

1.4 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 from three different projections. The first characterizes the aggregate access patterns of all users to all blogs. This is the blogosphere *server view* of the workload, which we present in Section 3. The second characterizes how individual users interact with the blogosphere. This is the blogosphere *user view* of the workload, which we present in Section 4. The third characterizes how individual blogosphere objects (blogs) are accessed. This is the blogosphere *object view* of the workload, which we present in Section 5. We put our work in context by reviewing related research in Section 6, and we conclude with a summary of findings and of current and future research in Section 7.

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 activities on the blogs by their owners.

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

Table 1: *Summary statistics of the blogosphere traces used in this paper* (*excluding requests by crawlers and requests that resulted in redirections or errors*)

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 read 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 blog-

gers 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 precise than an analysis that uses crawlers to characterize (say) the distribution of responses to blog postings, although we are not able to identify blog spams (postings or comments) using the information available on our workload.

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.

3. Blogosphere server view

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

3.1 Marginal distribution of transfer size: Figure 1 shows the complementary cumulative distribution function of the sizes of transfers from the blogosphere server. We have inferred that the distribution is heavy-tailed, namely a Pareto with parameter $\alpha \approx 2$. This result is greater than what has been observed for traditional web traffic [19, 6].

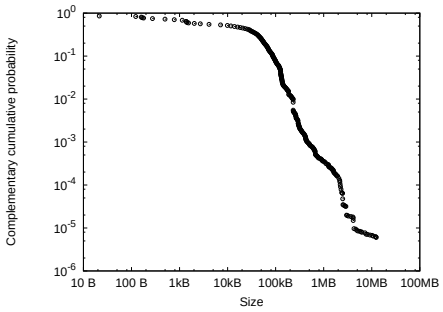

Fig. 1: *Marginal distribution (CCDF) of transfer sizes is heavy-tailed, namely a Pareto with parameter* $\alpha \approx 2$

3.2 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 2 illustrates this by showing the traffic 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 [18] 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.

3.3 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 quite smooth, which in turn result in low variability in the peaks observed in the diurnal access patterns (and mostly due to differences in intensity for weekdays versus weekends, for example [18]).[1] 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 3 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.

3.4 Traffic origin: In addition to characterizing our blogosphere traffic over time, we also looked at the origin of this traffic. In particular, we analyzed the origin (`referrer` field) of the various read requests in our traces and classified the traffic source as either *internal* or *external*. Requests from internal sources come through links from postings or comments in the same blog or a blog in the same blogosphere. Requests from external sources come through links in other blogospheres, social networks, and web sites, or links generated through search engines. Figure 4 shows the resulting breakdown of request origins.

4. 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 time since the last request in the session exceeds a timeout value (which we take to be 30 minutes).

[1] The only exceptions to this rule are web pages focusing on news and updates, whose popularity could change in a very short period of time due to external factors, such as crisis, wars, celebrity news, *etc.* We argue that for all practical purposes, such pages should be considered as "blogs". And, as we will show later, blogspace content focusing on news and updates represent a distinct class of blogs that are typically popular, but which do not underscore much blogosphere interactions.

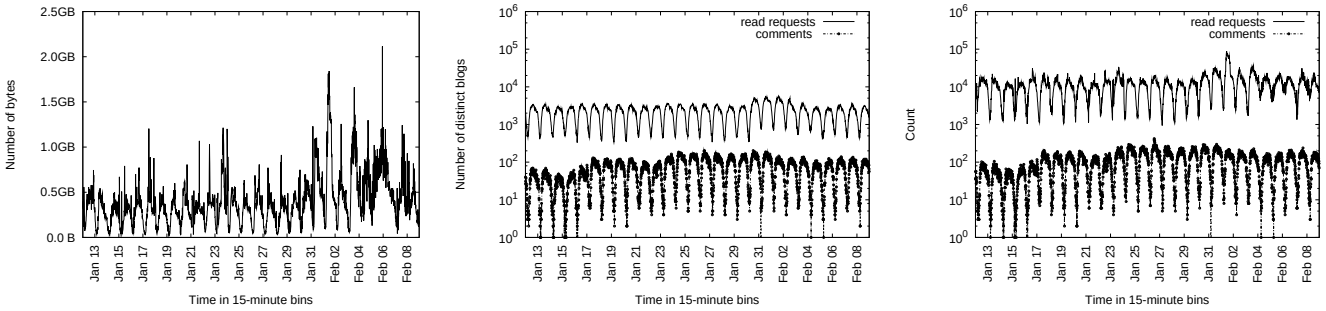

Fig. 2: *Diurnal Nature of blogosphere access patterns: Number of bytes transferred (left); number of distinct blogs accessed with read/write requests (middle); and number of read/write requests sent by blogosphere visitors (right)*

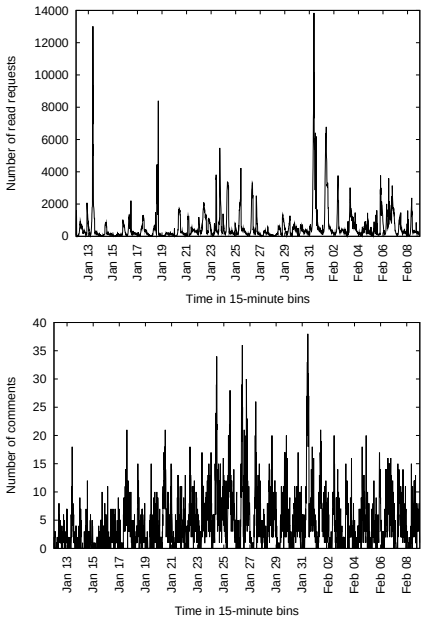

Fig. 3: *Diurnal access patterns underscore high variability in popularity of a given blog over time: Number of read requests (top) and comments (bottom) to the most popular blog*

4.1 Origin of user sessions: Figure 4 shows the fraction of user sessions (individual requests) that used search engines and web links to "enter" into our blogosphere (access its blogs). We obtained these figures by analyzing the referrer field of the first request of each session in our traces and each request therein. Of all the sessions in our logs, 29% did not have a value in the referrer field of their first request. These sessions were ignored in the results in Figure 4.

As evident from Figure 4, the largest chunk of traffic (and user sessions) into our blogosphere comes through search engines. Since search engines tend to rank their results based on popularity (*e.g.*, using page ranking algorithms based on web link structures), one would expect that popular blogs in a blogosphere would attract a disproportionate fraction of the traffic (sessions) emanating from search engines. To check if this is the case, we counted the number of requests that originated from search engines (as opposed to regular web links) and classified whether the requests were for the 5% most-popular blogs in our blogosphere (a total of 1,053 blogs) or for the remaining 95% of the blogs. Table 2 shows the resulting breakdown, which suggests (clearly) that search engines direct traffic more proportionally to less popular blogs as opposed to more popular blogs. This is an important observation because it suggests that in blogosphere the use of search engines has an egalitarian effect[14]. Another way of interpreting this result is that the popularity of the most accessed blogs is not the result of user "searching" for the blog, but rather it is the result of the user being "directed" to the blog through the link structures of (and the social networks underlying) the blogosphere.

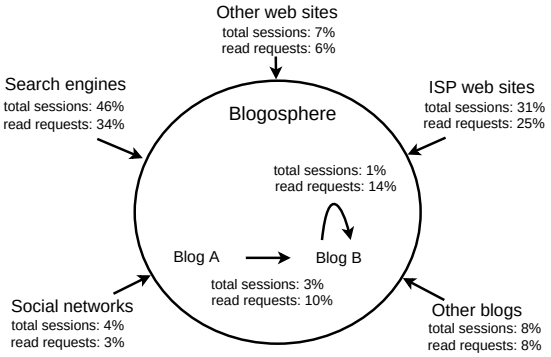

Fig. 4: *Fraction of sessions using search engines or external web links to "enter" the blogosphere*

Request origin	Fraction of requests to	
	5% most-popular	other 95% blogs
Search Engines	0.46	0.54
Web/Blog Links	0.63	0.37

Table 2: *Different ratios with which search engines and web links direct requests to popular blogs versus less popular blogs*

Figure 5 quantifies this observation further by showing the cumulative probability that a session originating from a search engine (versus a regular web or blogosphere link) will target a blog with a popularity higher than a given value (*i.e.*, a blog with a rank smaller than a given value). As evident from Figure 5, sessions originating from search engines are less likely to access highly popular blogs than those originating from regular web or blogosphere links.

4.2 User interest profile: 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 6 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 bl-

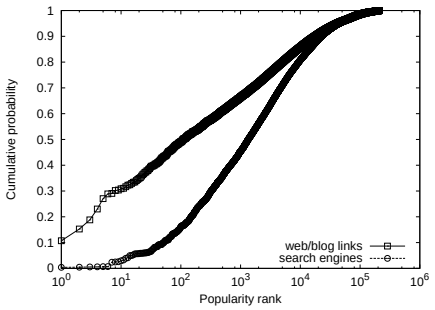

Fig. 5: *Cumulative probability that a session will target a blog with a popularity rank smaller than a given value, given that the session's origin is a search engine or a web link*

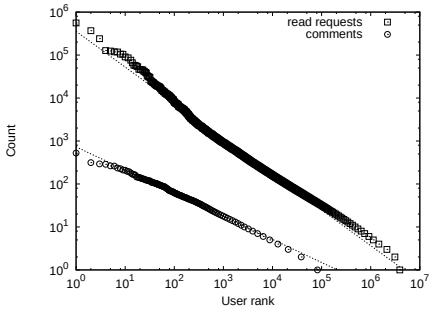

Fig. 6: *User access frequency versus rank of user interest*

ogosphere. The relationship in Figure 6 underscores a power law with $\alpha = 0.83$ (for read requests) and $\alpha = 0.54$ (for comments), both with $R^2 = 0.99$.

5. Blogosphere object view

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

5.1 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 blogger's level of activity: Does a high level of administrative activities for a blog imply a higher intensity of requests by visitors? Figure 7 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 7 shows that the correlation between the level of administrative activities and overall number of sessions accessing a blog is quite weak (if any).

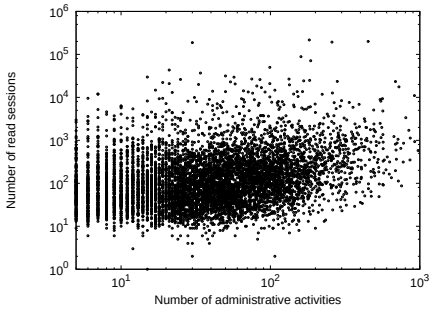

Fig. 7: *Total sessions versus number of admin request*

5.2 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 [7]. Thus, a natural question with respect to objects in blogspace (namely, blogs) is whether they exhibit a similar popularity profile. Figure 8 (left) shows the popularity profile of the blogs in our blogosphere by plotting on log-log scales the frequency of access (read and write) to a blog against the rank of the blog. The figure shows that 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. Figure 8 (left) shows that the popularity of objects in blogspace follows a general power law with skew parameter α. Using the total of read requests to a blog as indicative of popularity yields a skew of $\alpha = 0.97$ ($R^2 = 0.96$). Using total posted comments to a blog as indicative of popularity yields a smaller skew of $\alpha = 0.70$ ($R^2 = 0.97$). Figure 8 (middle) 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 8 (right) 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.

5.3 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 9. 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.

One set of attributes from Figure 9 that could be used to characterize the level of user engagement to a given blog is the interarrival time of user sessions, the interarrival time of postings, and the interarrival time of comments. We refer to these by the *inter-session*, *inter-posting*, and *inter-comment* times, respectively. Figure 10 shows the CCDF of the marginal distributions of these times. 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 11 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 9. 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 of the attributes we characterize in Figures 10 and 11, we show two distributions. The first is the aggregated distributions across all blogs, whereas the second is for the most popular blog in our blogosphere. In addition to the empirically observed distributions, Figures 10 and 11 also show the distributions that were best fits for our data. Table 3 shows these best fits.

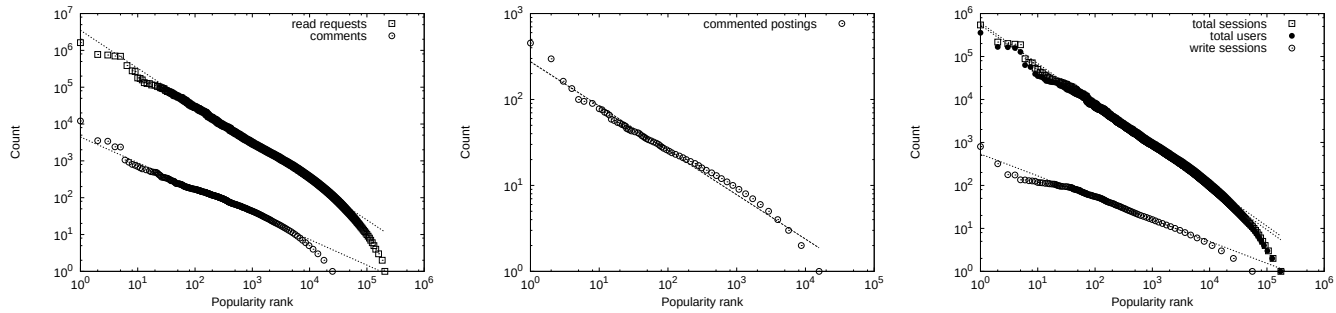

Fig. 8: *Popularity profile of blogs: Frequency of read/write requests (left); frequency of commented postings (middle); and frequency of total sessions, write sessions, and users (right) versus rank of blog. All profiles exhibit a power-law relationships*

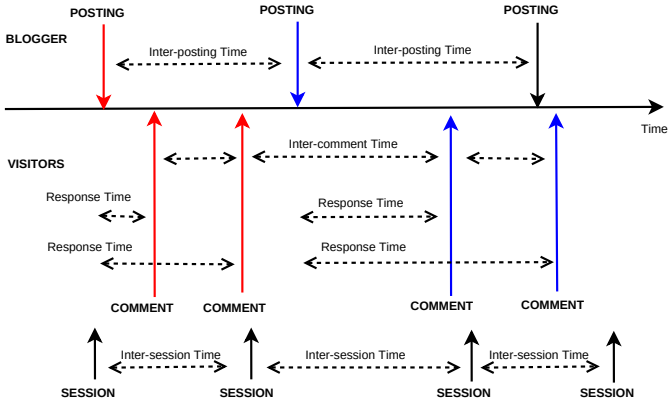

Fig. 9: *The dialogue structure induced by a given blog is defined by the actions of the blogger and visitors through the interleaving of postings with sessions and comments*

5.4 Classification of blogs based on interaction type: The above discussion makes the observation 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 blog.

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 comments-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 12 (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 great 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.

Interestingly, the results in Figure 12 (middle) 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 12 (middle) 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 12 (middle) 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 12 (middle) 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.

Based on the above observations, we grouped all blogs in our blogosphere into four categories based on their popularity and the ratio of write sessions they feature.[2] As illustrated in Figure 12 (right), *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 resulting breakdowns would be different for other blogospheres, but the basic observation (and methodology) would hold.

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

Interaction attribute	All blogs Distribution (parameters)		Most popular blog Distribution (parameters)	
Response Time	Weibull	($\alpha = 0.000469$, $\beta = 0.64892$)	Weibull	($\alpha = 0.000015$, $\beta = 1.04838$)
Inter-Session Time	Weibull	($\alpha = 0.069633$, $\beta = 0.33081$)	Lognormal	($\mu = 4.310535$, $\sigma = 1.40456$)
Inter-Posting Time	Gamma	($\alpha = 0.462894$, $\beta = 528,047$)	Gamma	($\alpha = 0.642546$, $\beta = 12,624$)
Inter-Comment Time	Gamma	($\alpha = 0.208459$, $\beta = 328,572$)	Lognormal	($\mu = 4.310535$, $\sigma = 1.40456$)

Table 3: *Distributions and associated parameters best fitted to the various interaction attributes observed empirically*

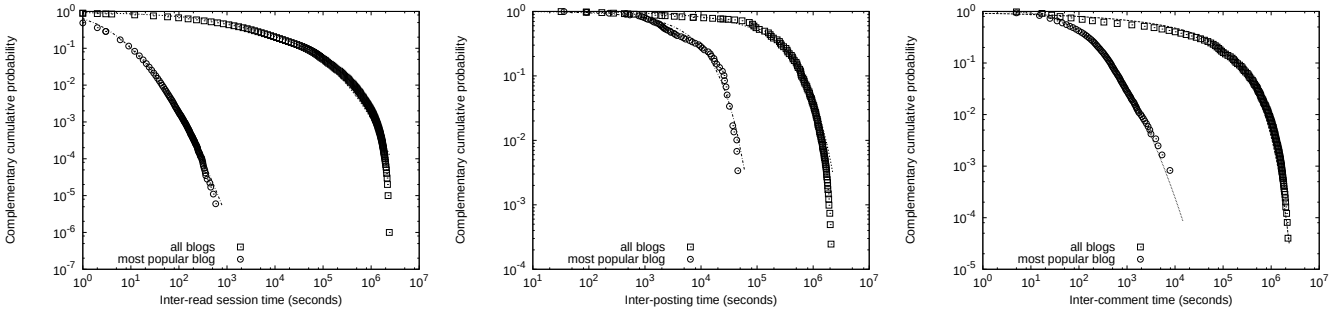

Fig. 10: *Distribution of inter-session time* (*left*)*, inter-posting time* (*middle*)*, and inter-comment time* (*right*)

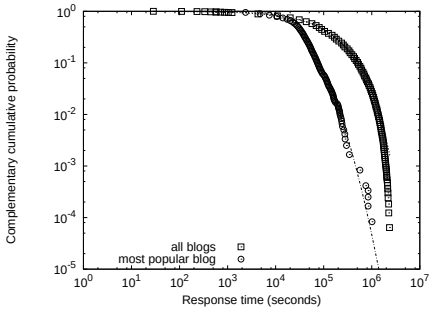

Fig. 11: *Response time distribution*

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%

Table 4: *Breakdown of blogs/sessions by interaction type*

6. 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 [17, 7, 2, 8, 3, 4]. 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 workloads involving traditional web content (while relevant to some aspects of our work) is outside the scope of this paper. 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.

There has been a number of studies that used blog and blogger characteristics for inference purposes. For example, Kolari, Java, and Finin [12] examined the "Splogosphere" of spam blogs (splogs) used to host spam postings. They provided a comparison of the characteristics of authentic blogs with splogs, which could be used to differentiate them. Herring *et al* showed that the interconnections between blogs can be used to characterize relationships between blogs, and to infer clusters of conversations and communities [10]. In [5], 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.

The results presented in this paper stand in some contrast to those observed in [15] 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 [13], 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 [1], 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 [9], 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.

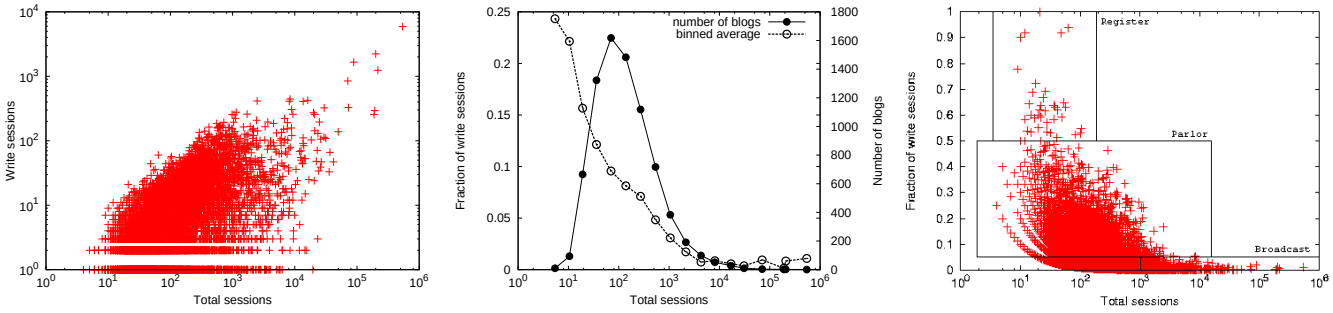

Fig. 12: *Interactions induced by a blog as reflected by total number of session and the fraction of write sessions: Scatter plot showing correlation between total number of sessions and total number of write sessions (left). Number of blogs and average ratio of write sessions for blogs with similar number of total sessions (middle). 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 (right)*

7. Conclusion

In this paper we used an extensive set of traces to characterize the access patterns in blogspace from three distinct perspectives: the blogosphere server, user, and object views. 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 conclusion that search engines have less of an impact on object popularity in blogspace, 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.

Our current and future work is focused on leveraging many of the findings and conclusions presented in this paper along a number of dimensions. First, we are looking into using the characteristics distilled in this paper as models to be used for the generation of synthetic blogosphere traffic. Given the increasing share of blogspace traffic, synthetic traffic generation is important for capacity planning and traffic engineering purposes. Second, our conclusion that the intensity of traffic directed to a blog through search engines (which use traditional page-rank algorithms) does not seem to correlate with the "real" popularity of the blog, suggests that social-network-based navigation may be playing an increasingly important role in web navigation in general, and blogosphere navigation in particular. On that count we note that in blogspace, the popularity of a blog is more a reflection of its owner's social attributes (*e.g.*, celebrity status, reputation, and public image) than a reflection of the number and rank of other blogs or web pages that point to it. This highlights the need for the development of page-rank algorithms that take into consideration the social attributes of blogosphere actors (as opposed to solely on the topology of the underlying blogspace). possibly using inference techniques such as those noted in Section 6.

Acknowledgments

The work of Azer Bestavros was supported in part by NSF awards #0524477, #0520166, and #0205294. Fernando Duarte was supported by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process #20060520221328a.

References

[1] E. Adar, L. Zhang, L. Adamic, and R. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, May 2004.

[2] V. Almeida, A. Bestavros, M. Crovella, and A. de Oliveira. Characterizing reference locality in the WWW. In *Proc. of PDIS*, December 1996.

[3] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in Web client access patterns: Characteristics and caching implications. *World Wide Web*, 2(1):15–28, 1999.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *Proc. of INFOCOM*, April 1999.

[5] E. Cohen and B. Krishnamurthy. A short walk in the blogistan. *Computer Networks*, 50(5):615–630, 2006.

[6] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Trans. on Networking*, 5(6), 1997.

[7] C. Cunha, A. Bestavros, and M. Crovella. Characteristics of WWW client-based traces. Technical Report 95-010, Computer Science Department, Boston University, 1995.

[8] Steven D. Gribble and Eric A. Brewer. System design issues for Internet middleware services: Deductions from a large client trace. In *Proc. of USITS*, December 1997.

[9] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. of 13th Intl. WWW Conference*, pages 491–501, 2004.

[10] S. Herring, I. Kouper, J. Paolillo, L. Scheidt, M. Tyworth, P. Welsch, E. Wright, and N. Yu. Conversations in the blogosphere: An analysis from the bottom up. In *Proc. of 38th Hawaii International Conf. on System Sciences*, 2005.

[11] The Blogger.com web site. http://www.blogger.com.

[12] P. Kolari, A. Java, and T. Finin. Characterizing the splogosphere. In *Workshop on the Weblogging Ecosystem*, May 2006.

[13] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of 12th Intl. WWW Conference*, pages 568–576, 2003.

[14] F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Googlearchy or Googlocracy? *IEEE Spectrum Online*, February, 1999.

[15] G. Mishne and N. Glance. Leave a replay: An analysis of weblog comments. In *Workshop on the Weblogging Ecosystem*, May 2006.

[16] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from a search engine perspective. In *Proc. of 13th Intl. WWW Conference*, pages 1–12, 2004.

[17] V. Paxson. Wide-area traffic: The failure of Poisson modeling. In *Proc. of SIGCOMM*, August 1994.

[18] E. Velloso, V. Almeida, W. Meira, A. Bestavros, and S. Jin. A hierarchical characterization of a live streaming media workload. *IEEE/ACM Trans. on Networking*, 14(1), 2006.

[19] A. Williams, M. Arlitt, C. Williamson, and K. Barker. Web workload characterization: Ten years later. In *Web Content Delivery*. Springer, 2005.

[20] The Technorati web site. http://www.technorati.com.