A Short Walk in the Blogistan

Edith Cohen, Balachander Krishnamurthy*

May 11, 2005

Abstract

The increasingly prominent new subset of Web pages, called 'blogs' differs from traditional Web pages both in characteristics and potential to applications. We explore three aspects of the *blogistan*: its overall scope and size, identification of emerging hot topics of discussion and link patterns, and implications both to blogs and applications such as search. Beyond blogs, we develop a general methodology of mining evolving networks and connections. The first part of our study is longitudinal— based on a five-week continuous fetch of a seed collection of nearly 10,000 blog URLs. The second part is based on a successive crawl of pages suspected to be blogs leading to a larger collection of several million URLs. The collection is examined for a variety of properties. We characterize blogs and study different facets of the link structure in blogs and its evolution over time, attributes of servers and domains that host many of the blogs including their IP addresses, and how blogs behave with respect to various HTTP/1.1 protocol issues. Inferences from our in-depth exploration are relevant to applications ranging from mining to hosting of blogs and other issues of relevance to the measurement community.

Keywords: Weblog; blog; hyperlinks; measurement; evolving networks

1 Introduction

What are blogs?

The word *blog* is short for the neologism "weblog", which is often a personal journal maintained on the Web. Blogs have grown rapidly in the last two years as a new communication mechanism between afi cionados who appear to avidly follow the opinions, stories, and observations. Blogs are similar in spirit to Usenet newsgroups except each newsgroup is a single person's view; some blogs allow for comments and a few blogs are shared between multiple authors. Often, a blog is one long Web page, partitioned into archives, with links to other URLs on the Web. In this sense, it is no different from a "home page" of a user. However, blogs in practice have turned out to be writings about a variety of topics, typically updated on a much more regular basis than homepages. Unlike homepages that are often maintained on individual sites owned by users, many popular blogs are on content hosting sites that provide space, software to maintain blogs, and generate indices, reverse pointer collections etc. Blogs are basically large queues (the term *blogroll* is used within the community) with additions appearing at the top of the page and older material scrolling down. Unlike a moderated site, additions to a blog is immediately available to anyone accessing the URL of the blog.

A typical blog consists of some text paragraphs often with embedded links (either internal links to another section of the same blog or external links), occasionally a few images, pointers to older sections of the same blog, and (in some cases) a set of reverse pointers to the blog itself made in other blogs. Many of the paragraphs (or blog sections) include a link (a paragraph-specifi c URL that others can use to refer to in

*The authors are with AT&T Labs–Research, New Jersey. E-mail: edith@research.att.com, bala@research.att.com

their blogs). While typical Web pages have a single point of entry (the URL), blogs have multiple locations of interest (the various paragraphs) and thus the link to specific paragraph has value.

Why are blogs worth examining?

Blogs are the fastest growing section of the World Wide Web in the last two years [1] and are emerging as an important communication mechanisms that is used by an increasing number of people. Although blogs began appearing several years ago they never crossed over to widespread popularity until 2000. By several estimations there are hundreds of thousands of blogs and as one might expect Zipfian in popularity and update frequency. Much like popular Web sites, blogs that are updated more regularly tend to be more popular.

Blogs are a distinct component of the Web from the viewpoint of content. There are several blogs that represent a small community of authors, i.e., content creators, with some communities (such as political blogs) that have a wide readership. The political blog `talkingpointsmemo.com` claims to have more than three hundred thousand unique readers in a month. The content creators routinely monitor related blogs and add links to items to related items on those blogs. This is done when there is a item in concert with views expressed, or when contrary views are discussed, or simply because it is relevant. This is one of the key differences between ordinary Web pages and a blog: the constant updating of content as well as links to other sites that are themselves changing.

There are several applications that can benefit from a characterization of blogs and we will discuss a few in this paper. Blogs offer a window into what many individual readers find interesting especially when new issues emerge. It has the potential for providing an early warning of hotspots and flash crowds. The Web site `slashdot.org` is an early example of a blog with significant impact on the future (and often short-lived) popularity of a particular Web site or Web page. Prior to popular blogs, often the source for hot news items were the news Websites. Unlike news sites, most popular blogs (with a few exceptions) are edited by a single individual. Many blogs allow anyone to comment on the contents. In fact the multiplicity of comments and additions of links to a new issue can be an early indicator of its rising popularity. A key distinction achieved by blogs is the original goal of making the Web a two- or multi-way medium rather than the widely prevalent "write once read many" model. Unlike news sites, popular blogs have the property of a large in-degree especially when one considers that links are not to the top-level URL but to a specific section of the blog. Blogs also offer interesting new collaborative filtering applications. Authors of blogs may be interested in finding out new stories that are related to stories they were previously interested in, blogs that are related to their blog, or ones that have commented on or linked to their blog. The last item is partly expressed through the publication of referrer links—a common blog phenomenon[1]

A blog can be a Web page or a site depending on how popular it is, where it is hosted etc. Although blogs change slowly, they are dynamic sites and represent the middle portion of the continuum between largely static sites (which are the vast majority on the Web) and the truly dynamic sites (sites that change regularly such as news sites). Search engines have distinguished between mostly static sites (home pages), dynamic sites (news etc.), truly dynamic sites (page generated upon visit each time, often ignored by search engines). The crawling, indexing, and search return phases of a search engine have taken appropriate action accordingly. Blogs create interesting new problems and opportunities in this regard.

[1] A referrer link is the url of the Web page from which a link was accessed, this information is often returned by the Web browser. Some blog pages extract and publish these links.

Overview

We refer to the blog space as the *blogistan* to describe the collection of blogs. Our contribution is threefold: We explore how emerging interests and patterns can be extracted by tracking a seed collection of blogs that have been modified fairly recently. By doing so we develop a methodology to identify emerging patterns on general data sets that comprise evolving communication networks. We examine the size and nature of the blogistan based on a recent collection of blogs. Finally, we present a collection of inferences and observations based on our study on identifying blogs, the growing spam problem in blogs, and how blog sites are accessed.

The rest of this paper is organized as follows: Section 2 characterizes blogs and discusses how they qualitatively differ from "traditional" Web sites. Section 3 describes the mechanics of our study and some key statistics related to it. Section 4 presents the analysis of the seed blog URL collection fetched repeatedly over a five week period in the autumn of 2003. We mine this data to identify emerging interests and patterns. Section 5 presents a walk through a large connected portion of the blogistan reached from our seed set. We examine the domain distribution of blog hosting sites and issues involving the HTTP protocol and blogs. Section 6 discusses inferences gleaned from our study with a preliminary analysis of Web server logs of a couple of very popular blog sites. We conclude with a look at work in progress on continued data gathering and analysis.

2 Differences between Web sites and blogs

There are several key differences between regular Web sites and blogs. Chief among them is that a blog is often a single page site; i.e., there are several related pages to the blog but found in archives and accessible from the main entry point page. The nature, number, and quality of links from a blog are quite different from ordinary Web pages. The primary reason for this is that blogs are often written to be read by many people, some of whom correspond with the blog authors to point out errors or related links. This allows the quality and richness of the blog to improve over time. Some blogs often consist mostly of contributed links (e.g., [2]). Many blogs, as stated earlier, are updated with significantly higher frequency than typical Web pages. The set of links originating from a blog are different from that of a typical page. As we will see later, a significant fraction of the links are to other blogs, thus constructing a close-knit community. Links to external sites that are not blogs typically are deeper links, since the text in a blog refers to a specific aspect that is covered in a page inside a site rather than to some top-level URL of a popular site. An example of this is the link to a specific news story under discussion at a particular time, which is found deep in a news site such as `cnn.com`.

Blogs are often personal journals or discussion groups on a narrow topic. Therefore, unlike pages on a large Web site, virtually the entire content of a blog is authored by a single person or a small group of people, leading to consistency of style, appearance, quality etc. Navigation through a blog is typically easier, since cross links to other blogs are a key feature of blogs. Additionally, given the few popular blogging software that are used heavily, there is considerable uniformity to a blog's appearance and a user's navigation experience.

Active (and the more interesting) blogs are updated with a frequency significantly higher than a traditional Web page (i.e., a home page of a single user). Often in a bursty manner. Inactive blogs will fairly soon notice a significant drop in accesses. We can use different measures for rate of change, amount of change, and last modification time in order to distinguish active blogs. Changes in blogs typically occur only at the top and thus it may be enough to fetch the first few hundred bytes via the HTTP/1.1 `Range` request or by using delta mechanisms [3, 4]). The number of new links added can be another metric but there are blogs that do not necessarily have many links but still have new text. Section 6.1 explores these issues in depth.

Traditional Web sites are designed as a coherent view of a subject, where older links may be as relevant as newer links. News sites, in contrast, are modified regularly with new content added while older content is archived away. On such sites, all content on the main page is new and expected to be relevant "now". Blog pages contain many old and some new links. The old ones are indexed by traditional search engines and are not relevant for the online discoveries. However, all links are explored in examining the blogistan.

3 Data gathering

We wanted to get a reasonable collection of Web logs to perform some characterization and measurement study. Since the Web has been around for over a decade there are several sites that rate popularity of Web sites. One reason for this is economical: Web sites used their ratings for computing rates for advertisement. Popular blogs have advertisement charges directly pegged to number of unique visitors and number of page views in a month [5]. The duration of popularity metric has allowed for some maturing of the sites that rate popular Web sites although their methodology is still somewhat murky. However, since the growth of blogs is a fairly recent phenomenon, the sites that rate blog popularity are few and their methodology is unverified. We decided that depending on popular blog ratings alone was not likely to be enough since our seed collection of blogs may not be representative of the blogistan. To offset this to the extent feasible we started with several hundred popular blogs based on a few blog popularity sites [6, 7, 8, 9, 10]) and added several thousand suspected blog URLs obtained from a list of URLs crawled in the spring of 2003 [11]. As a first cut any URL that had the string "blog" in it was deemed to be a candidate. However, we refined this list to eliminate duplicates, and obvious non-blog URLs. We examined the server portion of the URL string to ensure reasonable representation of various domains. Also, we eliminated blogs that had not changed within the preceding few weeks. We used the `Last-Modified` header timestamp to guide us. Our seed collection thus started with just over 10,000 URLs consisting of both popular and not-known-to-be popular blogs.

We decided to gather over a month's worth of data about these blogs fetching the URLs five times a day. In all, we had 171 instances of the seed collection of nearly 10,000 blog URLs gathered continuously between August 20 and September 23, 2003. For each of the blog URLs we obtained the meta-information (via `HEAD`), as well as the body (via `GET`). In the first phase of our study, we did *not* crawl the seed URL collection; i.e., we did not follow links. We ignored non-200 OK responses, javascripts, redirections, and other outliers. Gathering information multiple times over a contiguous period would allow us to examine changes in the contents, the link structure, as well as the rate of change. We ended up with a usable set of 8679 URLs for which we had predominantly `200 OK` responses, gathering a total of 171 times over a period of the 34 day study period.

For the second phase of our study, we extracted the links in each of the instances of each of the URLs both as a way of examining the individual blog page's link structure as well as an overall measure of how the blog collection differed from non-blog Web pages. Numerous studies have been done about the structure of typical Web pages and we expected the statistics about our blog collection to be different. The details of data gathering for exploring the size of the blogistan are discussed in Section 5.

4 Seed collection analysis

In this section we show how emerging interests and patterns can be identified by tracking our seed collection of the 8679 recently-changed Weblogs. We detect new referenced urls and study their emergence patterns. We also investigate to what extent standard tools, in particular hyperlink-based methods [12, 13] can be used to mine emerging new references from blogs. The first issue is the rate of change [14, 15] of blogs with

With dups.	Without dup.	URL
309	12	blogs.gotdotnet.com/tewald
262	11	www.xmldatabases.org/WK/blog
259	9	www.burtongroup.com/weblogs/annethomasmanes
258	8	linuxintegrators.com/blog/acoliver
98	7	www.diamondblog.com/archives/2003_09.html
98	7	www.diamondblog.com/archives/001577.html
98	7	www.diamondblog.com/archives/001576.html
98	7	www.diamondblog.com/archives/001574.html
98	7	www.diamondblog.com/archives/001571.html
98	7	www.diamondblog.com/archives/001570.html
95	4	www.diamondblog.com/archives/001575.html
77	5	www.blogdex.com
58	55	www.globalrichlist.com
45	40	www.democrats.org/blog
43	42	quizilla.com/users/jsimner/quizzes/How

Table 1: New urls with highest multiplicities (number of pages referencing them) before duplicate elimination, and their multiplicities before and after duplicate elimination.

respect to regular pages; most blogs do not change frequently. To account for changes, search engines deploy periodic crawls. Recent research has proposed adaptive incremental crawling techniques where pages that tend to change more often are crawled more often [16, 17]. These techniques allow for obtaining a less stale snapshot of pages. A second issue is that standard hyperlink-based methods analyze a *single snapshot* of each page. Therefore, even if this snapshot is the most recent one, there is no sufficient information of the time stamp of each reference. We argue that mining of blogs for emerging interests must incorporate the tracking of blogs over time, while placing a strong emphasis on newly added references.

As discussed in Section 3, we started with a seed set of 8679 suspected blog URLs and gathered 171 instances of these URLS. For each blog we then considered all URLs that were referenced in any snapshot of the blog. For each such blog-url pair we considered the first and last times of snapshots of the blog that included the url. The number of cases where a reference was introduced, dropped, and re-introduced were very few.

In our discussion we will use the term *new url*, for any url that was first referenced, by any blog in our collection, at least 24 hours past the measurement start time of noon August 20, 2003. We refer to other urls as *old urls*. These new urls are our candidates for having emerging interest. As a "data cleaning" measure, we identified different names for the same blog and referenced urls. For example, `blogmedic.blogspot.com:80/`, `blogmedic.blogspot.com/`, and `blogmedic.blogspot.com` all refer to the same page (this reduced the number of distinct blog urls to 8649). We limited our attention to new urls that had at least two references by different blogs in the measurement period. We obtained 4070 blogs that had at least one reference to such new urls, a total of 38903 urls embedded in these blogs, and 111876 unique (blog, new url) pairs.

To separate out the "interesting" new urls we applied some additional cleaning measures. As in traditional Web analysis, it is important to remove duplicate and near-duplicate pages so their references are not assigned higher weight. Since our emphasis is on *change*, we identify pages as near-duplicates if the new urls along with the time they were referenced are nearly identical.[2] Duplicate removal turned out to be very important due to the widespread use of *archives*—essentially copies of earlier versions of the blog

[2]In Section 6.2 we will argue that this methodology of comparing the change can be important as a measure against spam.

Multiplicity	URL
55	www.globalrichlist.com
42	quizilla.com/users/jsimner/quizzes/How
40	www.democrats.org/blog
36	www.nytimes.com/2003/09/18/opinion/18FRIE.html
34	quizilla.com/users/EerieFreek/quizzes/What
33	www.livejournal.com/users/beatings
33	www.bdmonkeys.net
33	bdmonkeys.net/~chaz/battle.php
32	www.cnn.com/2003/SHOWBIZ/Music/09/12/cash.obit/index.html
31	www.cjr.org/issues/2003/5/blog-welch.asp
30	www.editorandpublisher.com/ editorandpublisher/headlines/article_display.jsp?vnu_content_id=1979014
28	www.ritsumei.ac.jp/~akitaoka/saishin-e.html
27	www.usatoday.com/life/columnist/mediamix/2003-09-14-media-mix_x.htm
27	new.blogger.com/feature_giveaway/announcement.pyra
26	www.washingtonpost.com/wp-dyn/articles/A45450-2003Sep8.html
26	www.lileks.com/bleats/archive/03/0903/091103.html
24	www.thesmokinggun.com/archive/arnoldoui1.html
24	www.ospolitics.org
23	www.wumanity.info
23	www.tacobell.com/2003recall
23	www.foreignpolicy.com/story/story.php?storyID=13852
22	www.zogby.com/news/ReadNews.dbm?ID=732

Table 2: New urls with highest multiplicities (most references) after duplicates elimination.

page. These copies create a clique of pages referencing each other and sharing the same external references. Our 4070 blog collection had 25 urls as: www.kalilily.net:80/weblog/yy/mm/dd/nnnnnn.html, all archived copies of the same blog. These archives can be more explicit (with the term archive and date appearing in the url string) or implicit. In all cases we had observed, the archive shared the domain name, but simply unifying blogs with the same domain name is problematic, since there are blog hosting sites that host anywhere from a few to thousands of distinct blogs. For example, blogs.salon.com/nnnn are distinct blogs). In short, we had no uniform way to identify archives, but as we shall see, our duplicate elimination measure seemed to have countered their effects.

To remove duplicates, we identified all blogs with at least 3 new references where the first 15 new references along with times were identical (choosing 10 or 30 new references yielded nearly the same results). There were 1969 such blog clusters out of 2670 such blog urls. For all blogs with only 1 or 2 arrivals (time of observed first reference of a new url), we identified those with identical first and last reference times (1218 clusters out of 1400 blogs). We thus ended up with 3187 "distinct" blogs. Duplicate elimination turned out to be significant for distilling the "interesting" new urls based on the number of references made to them by distinct blogs (we refer to the number of such references as *multiplicity*): Table 1 shows the top referenced urls without duplicate elimination, the multiplicity before duplicate elimination, and the adjusted multiplicity. When contrasted with the final "cleaned" ranking in Table 2, we see that very large multiplicities were eliminated and that duplicate elimination significantly changed the "interest" ranking of the top new urls.

The next cleaning measures we applied is removal of blogs with an excessive number of new links. Since we are interested in blogs that are edited by a single or small number of humans, excessive number of new links is a sign that the bulk of references in the blog are less likely to be human edited. Some urls

in our set seemed to have thousands of new references. One reason for this was listing of `referer` links. Many of these links in pages we observed were search strings. Other "blog urls" with many new urls included blog-related sites such as `static.userland.com/weblogs`, `blogdex.media.mit.edu`, and `www.technorati.com`.
These sites automatically extract and list popular, recently changing, or otherwise interesting blogs. By considering the number of new urls in several known popular blogs, we chose 500 as a cutoff number. Applying this cutoff reduced the number of (blog, new url) pairs by a factor of two to 58825.

Table 2 shows the new urls with largest multiplicities in the cleaned data. These urls are referenced by distinct domains or blogs and seem to be related to news items or issues that were relevant at the data gathering time. Thus, we did not apply further cleaning measures. In Section 6.2 we discuss possible anti-spam measures that can be used if (and when) such data is a target of spamming.

4.1 Emergence pattern of references to new URLs

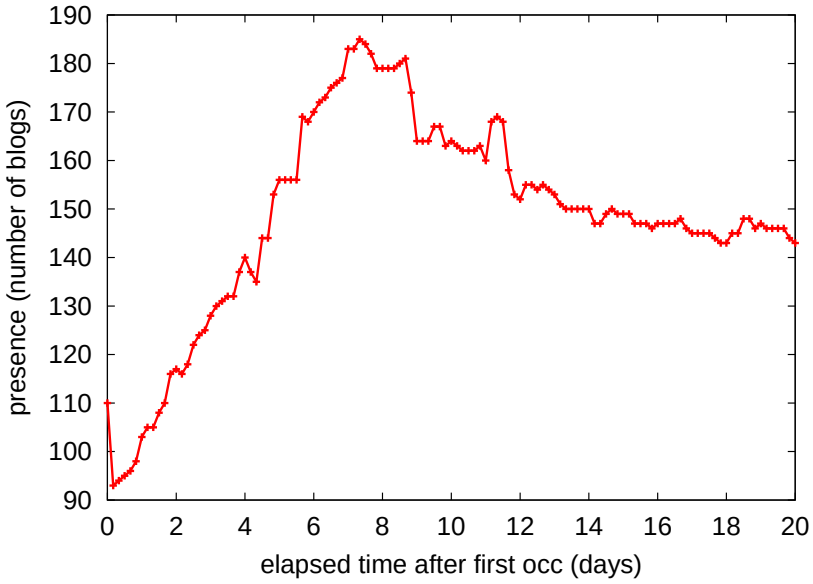
We next consider the emergence patterns of references made to new urls. We track the *presence*, that is, the current number of references to the new url, as a function of the elapsed time from the first time we recorded a reference to the new-url. Figure 1(A) shows the presence as a function of elapsed time, aggregated over all new urls that were first referenced during the first 12.5 days of the measurement period. Figure 1 (B) zooms on some new urls with high multiplicities that are selected from Table 2. (The data plots are cut off by measurement end time). The figure shows a clear bitonic pattern of an increase, peak, and decrease in presence. The rate of increase seems to be sharper than the rate of decline, which indicates that the lifetime of new-url reference typically extends beyond our measurement time frame. This is also explained by the scroll-down pattern in which blog pages are maintained, where new content is appended to the top of the page. The time to reach peak presence seems to vary from 1 to 12 days, with the "average" being around a week. This time presumably depends on the nature of the subject and how quickly it becomes stale. For example, the `tacobell.com` page that had to do with the September 2003 CA elections increased presence over a longer time period. The CNN Johnny Cash obituary, in contrast, quickly reached its peak presence.

4.2 Mining for emerging new urls

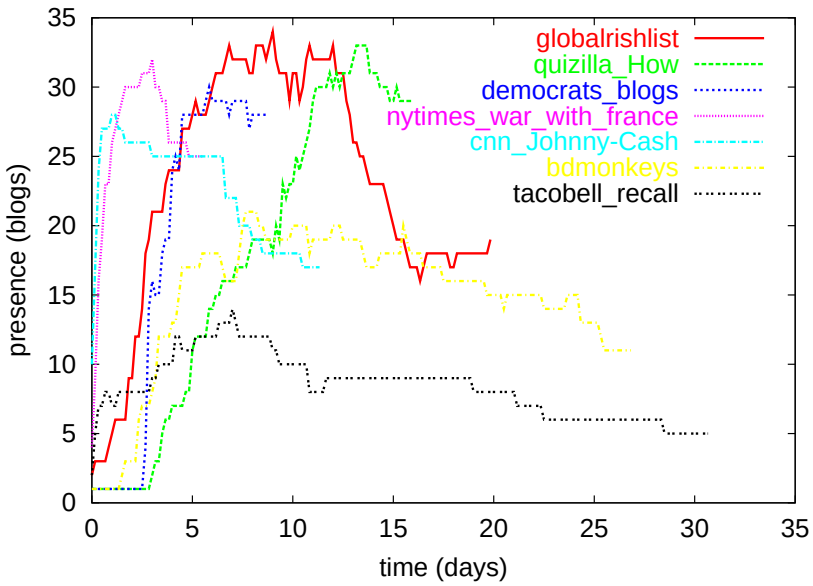
Blog pages contain many old and some new links. The old ones are less interesting as they are indexed by traditional search engines, and are not relevant for the online discoveries we seek. To distinguish properly, one needs to compare different snapshots of the same page (i.e., incorporate the time dimension in the analysis). Because of their basic patterns, analysis of single snapshots can work well both for static Web pages and for news sites but not for blogs.

Traditional, i.e., static hyperlink analysis depends on in-degrees. Roughly speaking, importance of a page depends on the amount and quality of references to it. For analyzing blogs, the urls we care about are the new ones. But at the point we are interested in mining them they have considerably fewer references than other older urls. Figure 2 reveals an order of magnitude gap between the multiplicities of old urls and those of new urls (note that 'old' and 'new' add up to 'all'). In our blog collection, the top referenced urls overall had thousands of references, with the two top ones being `www.blogger.com` with 5134 and `www.blogspot.com` with 4484 references. In contrast, the top referenced new-urls had only few dozen references (see Table 2).

The upper line in Figure 4 shows the distribution of the ratio of new-urls to total number of references per blog page. The lower line of the figure plots the ratio of *new references* to all references in a blog, where *new references* are defined as links that are new to the blog, in that the first snapshot of the blog they appeared in was at least 24 hours past the start of the measurement period. Note that new references



(A)



(B)

Figure 1: (A) Presence (# of references) as a function of elapsed time from the first observed reference, aggregated over all new urls appearing in the first 12.5 days of our measurement period. (B) Presence as a function of elapsed time for selected, high-multiplicity, new urls.

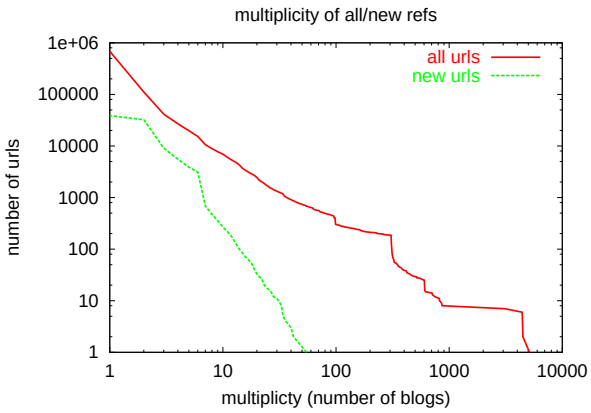


Figure 2: Distribution of the multiplicity of all urls and of new urls.

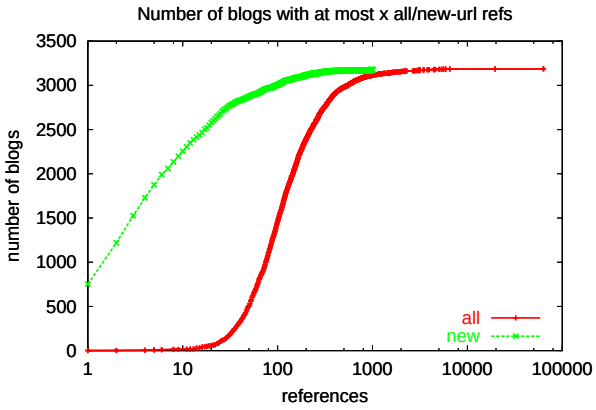


Figure 3: Blog size distribution in terms of total number of references and total number of new-url references.

are not necessarily new-url references as references to these urls possibly existed in other blogs at the start of the measurement period. This distribution is indicative of the "amount of change" of blogs during the measurement period.

Figure 3 shows the distribution of "blog size" in terms of all references and new-url references the blog contained in the measurement period. We can observe that the size distribution in terms of all references has the form of a bell curve which peaks at several hundred references. For new-url references, the dependence is monotonically decreasing with more blogs experiencing fewer changes.

These results show that new-url references are vastly outnumbered by the sheer number of other references over a collection of pages. Moreover, even on a per-page basis, new-url references are outnumbered by other references (that is, new-urls references do not even dominate references in pages where they tend to occur). Furthermore, the number of new references made to new-urls is considerably smaller than the number of new references made to old urls. This implies that identifying emerging new urls requires tracking many pages, over time, and combining the information; as they can not be distilled from a single snapshot of many pages nor from tracking individual pages.

When contrasted with Figure 1(A), the data for new urls in Figure 2 provides an indication at what presence level we can start "paying attention" to a url. There were about two hundred new urls with at least 10 references, which means that some of the "slow" rising emerging new urls can go unnoticed for several

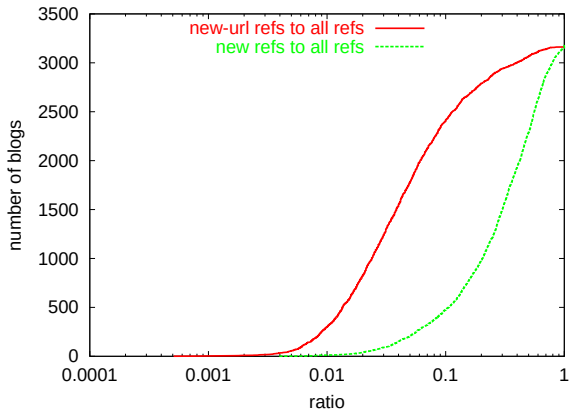


Figure 4: Cumulative distributions of the ratio of new-url references to all refs per blog and the ratio of the number of new references to all references per blog. The figure plots the number of blogs with ratio at least x.

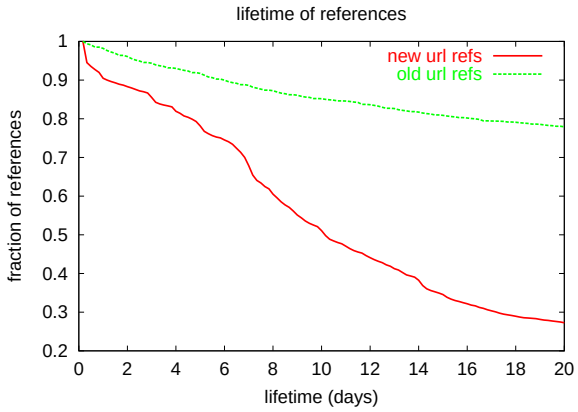


Figure 5: Lifetime durations of new-url and old-url references

.

days.

Figure 5 shows the distribution of lifespans of references to new and old urls. There were 18.5K references to new urls in our cleaned data set that first occurred in a snapshot taken during the first 12.5 days of the measurement period. There were 305K references to old-urls made by the same set of blog pages. These references were tracked for 20 days following their first appearance; we have full remaining 20 days of snapshots to identify the time these references were removed past the first 12.5 days. The figure shows that new-urls references are much more transient than old-url references. Half of the new-url references lasted for at most 10 days and only a quarter lasted beyond 20 days, whereas 80% of old-url references stayed put for the 20 days. Therefore, the recognized pattern where objects that had not changed recently are less likely to change appears to hold to references as well.

A recent paper [18] also recognized the importance of time of the appearance of a reference. Their focus is on detecting 'communities' (group of blogs that have a burst of inter-references). We focus on general properties of the blog graph obtained via a crawl, and specifically on using blogs as a tool to identify interesting emerging topics (refs) that are mostly external to the Blogistan. Their methodology for recognizing link 'age' is different: while sufficing as a proof of concept analysis, it is less general. They use a one-time

crawl of blogs from 7 specific sites and rely on heuristics based on archives to determine which links are new. To recognize archives, they use a simple heuristic of looking for the string "Archive" in the URL. We observed that in many cases this is not sufficient. Also this may require last-modified time stamps on archive files which are often unavailable. Finally their approach may not capture finer time granularities as compared to ours where we fetched data every 4 hours.

5 The blogistan

Moving beyond our seed collection, we wanted to examine a larger fraction of the available set of blogs—the blogistan. A few studies have reported the size of the blogistan to range from 1 Million to over 4 Million. A study [19] (based on a collection of about 3500 blogs) reported that a significant fraction of blogs—about two thirds—had not been updated in two months. However, this study only used blogs maintained on certain blog-hosting services. Of the 375 popular blogs in our seed set several are not hosted by such sites.

The 8679 blog URL seed collection fetched repeatedly over 34 days yielded over a million (1.138 Million) URL links emanating from them. Nearly a fifth of these URLs (223K) appeared to be blogs. As a preliminary measure of getting more information about the blogistan, we obtained a *single* copy of the contents (and meta-information) of the 223K blogs linked from the seed collection. Note that any URL that had the case-independent string 'blog' was considered to be a blog. While some blogs may be missed as a result and some non-blogs may have been included, our subsequent fetches significantly reduce any impact of this.

Of this 223K first-hop blog collection, 165420 (74%) of the responses had `Last-Modified` response header. Of these 72542 (43.8%) had been modified at least once in October 2003 (the month they were fetched). 33.5% of the blogs had *not* changed in two months. This 223K blog collection is a significantly larger sample than the seed collection. We extracted the links from each of these 223K pages to examine if they were blogs, and obtained a distribution of the counts of links etc.

The second step in fetching the 223K blogs (`HEAD` and `GET`) resulting in 7.65 Million unique pages. Of these, 1.7 Million unique pages were identified as blogs (again by looking at the URL string). Among the 978794 HTML URLs (non-images etc.) there were 467168 (47.7%) responses with `Last-Modified` headers. Of these, 219911 (47.07%) had been updated at least once in October 2003, 170814 (36.56%) had *not* been changed in two months.

In the next stage of the crawl we eliminated blog URLs already seen and fetched 823438 new URLs. This resulted in 5.325 Million pages of which 1.46M were identified as blogs. The diminishing number of new blogs at each hop of the crawl is an indicator that we are converging on obtaining a large fraction of the connected portion of the blogistan reachable from our seed set. While it is possible that there may be significant sections of the blog space that we have not explored, given that we followed most blog links, the probability is likely to be low. Of the 823438 pages fetched, 273337 had Last-Modified response headers of which 140699 (51.47%) had been updated at least once in October 2003, and 106209 (38.86%) had *not* been changed in two months. There is thus a confirmation that roughly more than a third of the blog space was not being changed actively. The crawl at this level showed a even smaller number of new blogs to fetch and we decided to stop our crawl at this level to get a first measure of the blogistan. We plan to continue the crawl. In our three steps of the crawl, a blog page had an average of 48 unique links.

We refer to each crawled URL as a *blog page* (with blog URL string) and each outgoing link as a *reference URL*. With a collection of blogs at several steps, we realized that many of the links in our collection were duplicates—the identification of "duplicate" pages are in terms of having identical or near-identical set of reference links. Our duplicate elimination process involved case homogenizing the URL strings, removing trailing back slashes, ":80" after the domain names, and trailing #∗ (referring to different places on the same page). If the domain names of the blog and reference URL were identical, then the #∗ was

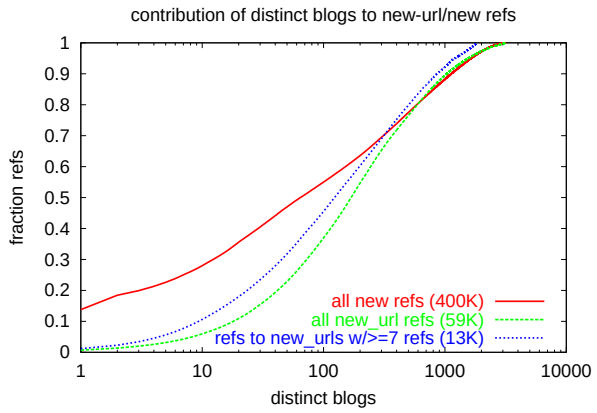


Figure 6: Cumulative fraction of blogs with at most x distinct blog URLs, domains, and second-level domains.

removed also from the reference URL. This is because archiving software "translates" internal references to the new page, making it seem like non-identical copies. We carried this out for the union of URLs at all the hops of our crawl and ended up with 1,239,827 (roughly 1.24 Million) distinct blog URL strings.

We then looked at a random hash value for each referenced URL and created a "signature" for each page that constitutes of the (ordered) 10 keys with minimum hash value. Blog pages with less than 10 references had a signature that exactly captured these references. This is a fairly standard use of the min-hash technique[20]. Such a technique has been used for text-based duplicate elimination [21]. We obtained 749775 unique signatures which correspond to the same number of clusters of blog URLs. This is roughly a 40% reduction in the 1.24 Million distinct blog URLs.

Figure 6 shows the distribution of the number of distinct blog URLs, domains, and second-level domains, per cluster. The fi gure shows that the similarity was heavily biased towards blog URLs that share a domain and more so a second-level domain. While 10% of clusters had more than one blog URL, less than 1% had more than one second-level domain.

5.1 Blog domains and related attributes

We examined the distribution of domains where blogs are hosted. Unlike traditional Web pages where most of the pages are hosted in a relatively distributed fashion, there are thousands of blogs including several popular ones that are hosted in a few blog-hosting sites. Although there are popular several eponymous blogs (e.g., `lessig.com/blog`), some popular blogs are hosted anonymously under opaque names. For example, the highly popular Julie/Julia project's blog URL was hosted under `salon.blog.com` as `salon.blog.com/000139.html`).

There were 234524 distinct domain names in our collection. Looking at the (lowest in a sorted order) domain name from each cluster showed that there were 182665 unique names. This is due to domain aliases that point to the same pages. One example is (`literatus.blogspot.com`,`www.literatus.blogspot.com`). As for second-level domains, there were 14528 second-level names among the unique representatives of each cluster (15081 among all blog URLs). The large gap between the number of domain and second-level domain names is partly explained by aliases and some popular blog sites. For example, `blogspot.com`, assigned a unique domain name for each blog, (e.g., `sykur.blogspot.com`, `rotolalavita.blogspot.com`, `dacouver.blogspot.com`). Figure 7 shows the number of unique blog pages per domain (out of 234524 domains) and per second level domain (out of 14528 names).

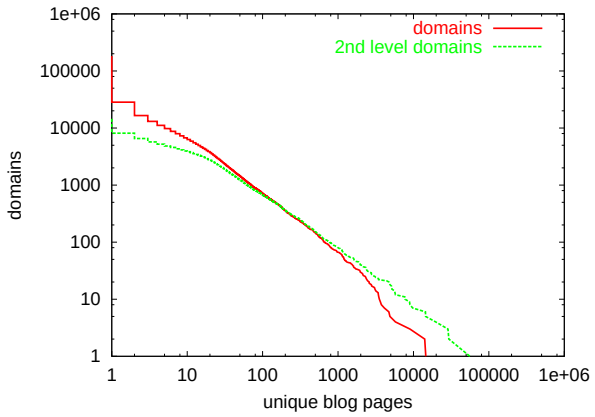


Figure 7: Cumulative number of domains, and second level domains, with at most x unique blog pages.

However, many of these 'unique' names are just aliases for the same IP addresses. A DNS lookup of all the 234,524 distinct domain names showed that they mapped to just 11,870 unique IP addresses. These 11,870 IP addresses were further clustered into just 3321 BGP network-aware clusters [22]. Within the set of clusters, more than 60% contained single address indicating the broad distribution across different prefi xes. However, a relatively small concentration in 11,870 addresses can make some blogs targets for DoS attacks. A report [23] in October 2003 discussed a denial of service attack on a set of blogs that were discussing a specifi c topic. Such incidents are likely to increase and at least the popular blogs would have to be replicated using CDNs or other such technology to protect them. A few thousand IP addresses hosting most of the blogs represent a surprising concentration in a relatively well-known and avidly followed sub-culture on the Web.

5.2 Blogs and HTTP protocol issues

Among the complete collection of blogs 78.4% were hosted on Web sites running the Apache Web server indicating the continuing popularity of Apache. Of the remaining server types, IIS accounted for 15.3%. The `Last-Modified` response header, a crucial header for checking staleness was only present between 35% to 48% of the blog URLs in the different crawl hops. The most popular protocol variant is HTTP/1.1. We did not fully test protocol compliance [24] in this study. Instead, we examined HTTP protocol factors that might be of particular relevance to blogs: certain requests, the ability to handle certain request headers, and presence or absence of response headers.

In general blogs are poor choice for caching since they change a lot; however, as we have discussed earlier, they change in a peculiar way. Unlike typical Web pages, new material is added to the top of the blog, while with lower frequency, old material is removed from the bottom. In order to get a fresh and semantically meaningful copy of a blog it should be enough to get the new bytes and occasional full copy fetches to sync up the removed portion. Thus, the ability to handle `Range` requests or handle delta updates [3, 4] can be crucial. Also, the use of certain HTTP/1.1 request headers such as `If-Unmodified-Since` can reduce the number of bytes transferred.

Compressing blog pages might be useful since they are mostly text. However, from the viewpoint of delta, the average number of new bytes added is low. When tested on hundred random seed URLs in our collection across the 171 instances over the study period, 30% of the URLs changed on average by less than 100 bytes, more than 60% of the URLs changed on average by less than 200 bytes, and 98% of the URLs changed by less than 1000 bytes. The usefulness of `Range` requests is further validated upon the examination of the content length of these 100 blogs. Around 20% of these blogs are less than 10K bytes

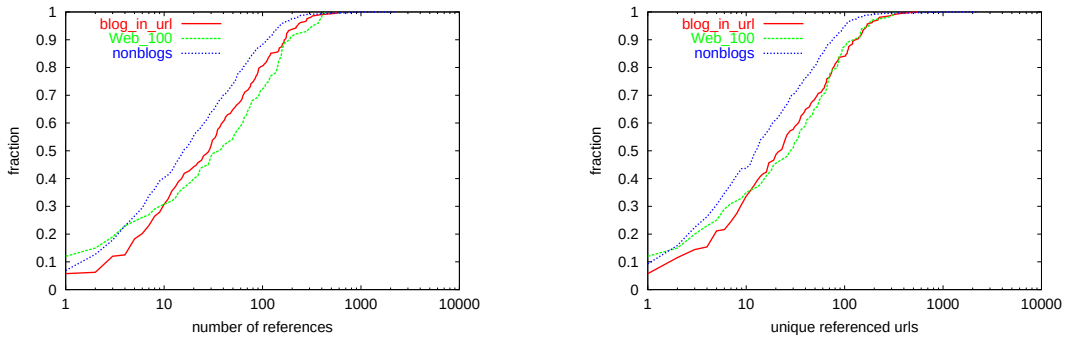


Figure 8: The number of links (left) and unique links (right) in the three datasets

and nearly two-thirds are less than 40K bytes and the remaining third are between 40K and 80K bytes. A range request of 1000 bytes would significantly reduce the traffic for popular blogs. Unfortunately, a test of nearly 2000 random blog servers showed that only around 40% are able to handle HTTP/1.1 Range requests successfully returning the `206 Partial Content` response. We tested this by sending `GET` requests with byte range 1-10 using *httperf* [25]. Oddly enough, a handful of servers returned the invalid response of `416 Requested Range Not Satisfiable`. We also tested how many of the blogs handled a `GET` request properly in the presence of `If-Modified-Since` header and return a `304 Not Modified` header. We sent requests successively with a current time stamp and yet only 30% of the nearly 3000 random blog URLs tested gave the appropriate `304 Not Modified` response. Only about 10% of the 3000 responses had `Expires` header and nearly half of these were in the past. Very few of the blogs use `Content-Encoding` or `Expires` header. The most popular language when the `Content-Language` was present was not too surprisingly English.

6 Inferences from our analysis

Thus far, we have presented the results of our study comprising of analyzing a seed collection of blogs, crawling to get a large fraction of the blogistan, and the subsequent protocol-level analysis. Below we present observations gleaned from our study. The study can help us provide a clear idea of what a blog looks like. Automatic identification of a Web page as a blog can help in modifying crawler algorithms, smarter indexing, and distilling search results. Identifying sub-communities of related topics automatically would be enabled. Browsers and other blog related software can use our analysis to improve fetching of blog pages. We examine the current manner and extent of spamming in blogs. By examining a couple of server logs belonging to popular blogs, we can get a view of how blogs are linked and accessed. We stress that these are preliminary observations based on our study.

6.1 Identifying a blog

Our heuristics for checking if a Web site is a blog or not involve various tests whose results taken together can indicate with high probability if a specific Web site is likely to be a blog. The steps include examination of the number of links in the page: blog pages rarely have a count of zero links. The higher the number of links the more likely the page is a blog. Examining the nature of links shows that often there are several links to the blog URL itself (either to facilitate reverse linking by others or to refer to another part of the same blog). Blogs tend to refer to other blogs—our preliminary analysis shows that about a quarter of all links from a blog are to other blogs. The presence of certain terms (e.g. RSS, blog, xml) in the body is a

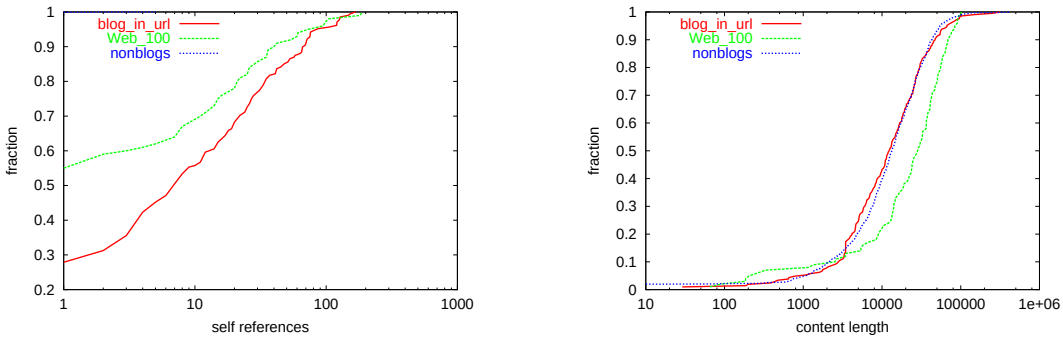


Figure 9: The number of self-reference (left) and content length (right) in the three datasets

positive indicator of a blog. The occurrence of the string blog/BLOG in the URL is another such indicator. Many bloggers go out of the way to identify their pages as a blog rather than try to obscure it. The content length of blogs is likely to have a different characteristic than regular Web pages.

To validate our hypothesis we conducted a series of tests: we selected three sets of URLs. The first set 'bloginurl' is a set of 226 randomly chosen URLs from our seed collection that have the string 'blog' in them, also known to be blogs. The second set 'web100' consists of 100 popular [26] Web pages that may or may not be blogs (in fact none of them are likely to be blogs). The third set 'notblogs' is 226 randomly chosen Web pages from the 154 Million URL crawl collection [11] with which we began our study. This set is not likely to have any overall popular Web pages or a blog. We fetched the meta-information and contents of these three sets and examined their content lengths, the number of links emanating from them, the number of unique such links, and the number of *self-references* (hyperlinks pointing to points on the same Web page). Figure 8– 9 shows the CDFs of each of these attributes for the three datasets.

As Figure 8(a) shows, popular Web sites and blogs have considerably more references than unpopular Web sites that are not blogs. But popular Web sites generally appear to have more references than blogs. Examining unique references shows (Figure 8(b)) that this difference between popular Web sites and blogs disappears while nonblogs continue to have a lot fewer unique links. As we can see from Figure 9(a), blogs have significantly more self-references than popular Web pages. Unpopular Web pages have practically insignificant number of self-references. Self-references are thus a key indicator of difference between blogs and Web pages. In terms of content length (Figure 9(b)) that blogs appear to be smaller than the popular Web pages while non-blog (unpopular) pages appear to be similar to blogs. Note that these are based on single point snapshot; the nature of links emanating from blogs (with a significant fraction going to other blogs) is another key difference.

6.2 Anti-spam measures

Our notion of blog spam is similar to the one used in the context of search engines: Artificially modifying behaviors, replicating pages, inserting hyperlinks, or modifying referrer fields in order to be disruptive or attract undue attention.

By displaying the content of the `Referer` field of incoming requests, blogs inadvertently provide free space for spammers. Since the referer field can be altered to anything (hand crafted browser or a simple script that generates the request), a spammer can place any links of their choice on a large number of blog pages. This is done to boost their indegree and ranking in some search engine rankings. Some search engines have asked blog hosting sites [27] to route their links through them so they can ignore such inflation. Even without spamming, referer displays add large numbers of relatively meaningless link references. Additionally, many blogs provide space to post comments on a particular issue. This mechanism also seems to be hijacked

to some degree by spammers. One way to deal with this is to come up with automatic means (possibly by analyzing popular blog authoring tools) to distinguish such references from true authored references. Another is to simply ignore blogs with many new urls. The archives also artificially increase the number of references to a link but our duplicate elimination techniques can handle this.

Intentional spam, which we have not noticed in our study, but is known to exist, can amount to creating large numbers of fake blog pages and placing arbitrary content on them. Some of this content may be considered high quality with references to established and related sites and then mixed in with particular spam links. Such spam can be dealt with by tracking blogs over time and building a profile and authority weight for each blog. Newly discovered or created blog pages as well as blogs that tend to have large number of new links can be assigned lower weight. Blogs that seem to be focused (pointing to related sites) and pointing to new urls that in retrospect turn out to be popular are considered more authoritative. Thus, popular political blogs, which tend to raise issues that are quickly followed up on, will obtain higher authority weights, and so will other blogs that tend to point to issues pointed to by those blogs.

6.3 Preliminary analysis of server logs of popular blogs

Until now we have only examined the contents of blogs in terms of links and their meta information. We now turn our attention to how a blog site views accesses and for this we need web server logs of blogs sites to get a better idea of use of these blogs. We were able to obtain a couple of Web server logs of two very popular blogs. Both these blogs appear in various popular blog listings (e.g., Bloglines [28]). Due to privacy reasons we are not disclosing their names. The first log, H, received 408579 requests from 77888 unique IP addresses during the same 34 day seed URL collection period (August-September 2003). The other log, G, received 963565 requests from 33454 unique IP addresses in a two week period in late October 2003. We were able to obtain this log only for this time period. Although these are popular sites, our examination of just two server logs is *not* intended to be representative. We are in the process of obtaining more server logs of popular blogs. However, examining even these two popular blog sites can offer some interesting clues to blog access.

In both logs, the most common request was to the top level blog URL (74% in H and 38% in G). Close to half of the accesses were for different images (banners etc.) in G. We examined the referrer field (present around 30% of the time in H and 56% in G) in the server logs to see both where most of the accesses were originating. We were also interested in seeing if the accesses were to archived sections of the blogs. Around 15% of the references in H were from the blog site itself (i.e., users moving from one section of the blog to another) while it was much higher for G indicating that people spent more time navigating within the G site or fetching the images. Nearly a third of external site reference were from a single site for H. Similarly a third of external site reference to G were also from a single site: a news aggregator site like Radio Userland [29] which reads a variety of news sources and shows recent postings on one page. We examined the behavior of a popular search engine crawler on both sites and there appeared to be no discernible distinction made due to the site being a blog when compared to other non-blog URLs. We should stress that this is a preliminary measure and we will be able to better report when we have a dozen or so server logs of popular blogs. Blog-specific crawlers (e.g., MIT Media' Lab's Blogdex [30], Daypop [31]) have been written to circumvent the problem of having to rely on large diameter of search engine spiders.

7 Conclusion and ongoing work

Given the popularity of blogs in popular culture [32] and their rapid growth as a distinctive part of the Web, it is natural to examine this phenomenon. Blogs provide a multi-way communication paradigm on the Web that typical Web pages do not. The rate of change of blogs is quite different from traditional Web pages

and the nature and count of links between blogs and other Web pages are quite distinct. The rich content reflecting the human authoring and the steady updating indicates continuing interest on the part of the large number of readers and other bloggers.

Ours is the first widespread study to characterize individual blogs and the shape and size of the blogistan. We have characterized the key differences between a Web page and a blog. We have shown how to identify emerging interests and argued that in contrast to traditional search, this mandates not only large scope analysis of referenced links but also of their evolution over time. Our multi-hop crawl has presented a reasonable approximation of the connected portion of the blogistan along with several of its interesting attributes such as the number of unique domains that host blogs and the relatively low number of IP addresses. We have examined which HTTP protocol issues are relevant in dealing with the blogistan. We have started our analysis of Web server logs belonging to a couple of popular blogs. We point out the risks due to spamming and how our duplicate elimination can help in a variety of ways.

An important contribution of our work is the methodology we developed to identify emerging interests by mining hyperlinks in blogs and their change over time. The methodology constitutes a general approach to mine evolving interconnection networks that we believe can have applications well beyond the Blogistan. By canceling out ”repeated patterns” we are able to identify emerging ones. One example applications from a different realm is to study ISP level Netflow data over time and identify distributed bot attacks on particular hosts, detect new Internet worms, and predict flash crowds.

References

[1] R. Blood, “weblogs: a history and perspective.” `http://www.rebeccablood.net/essays/weblog_history.html`.

[2] D. Barry, “The (Un)official Dave Barry Blog.” `http://davebarry.blogspot.com`.

[3] J. C. Mogul, F. Douglis, A. Feldmann, and B. Krishnamurthy, “Potential benefits of delta encoding and data compression for HTTP,” in *Proc. ACM SIGCOMM*, pp. 181–194, Aug. 1997. `http://www.acm.org/sigcomm/sigcomm97/papers/p156.html`.

[4] J. Mogul, B. Krishnamurthy, F. Douglis, A. Feldmann, Y. Goland, A. van Hoff, and D. Hellerstein, “Delta encoding in HTTP,” RFC 3229, IETF, January 2002. Proposed Standard `http://www.ietf.org/rfc/rfc3229.txt`.

[5] “Blogads for opinion makers.” `http://www.blogads.com/order_html`.

[6] “Salon radio community server.” `http://blogs.salon.com/rankings.html`.

[7] “Top 100 Technorati.” `http://www.technorati.com/cosmos/top100.html`.

[8] “Most watched blogs.” `http://blo.gs/most-watched.php`.

[9] “The blogosphere power rankings–the most popular political blogs on the net.” `http://www.rightwingnews.com/special/topblogs.php`.

[10] "Userland site report."
http://stats.userland.com/groups/radio1/report.html.

[11] M. N. Dennis Fetterly, Mark Manasse and J. Wiener, "A large-scale study of the evolution of web pages," *Software Practice and Experience*, May 2004.

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th World Wide Web Conference*, 1998.

[13] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[14] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul, "Rate of Change and other Metrics: A Live Study of the World Wide Web," in *Proc. USENIX Symposium on Internet Technologies and Systems*, pp. 147–158, Dec. 1997.
http://www.research.att.com/~bala/papers/roc-usits97.ps.gz.

[15] B. Brewington and G. Cybenko, "How dynamic is the web?," in *Proceedings of the 9th World Wide Web Conference*, 2000.

[16] J. Cho and H. Garcia-Molina, "The evolution of the web and implications for an incremental crawler," in *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pp. 200–209, Morgan Kaufmann, 2000.

[17] J. Edwards, K. McCurley, and J. Tomlin, "An adaptive model for optimizing performance of an incremental web crawler," in *Proceedings of the 10th World Wide Web Conference*, 2001.

[18] R. Kumar, J. Novak, and P. Raghavan, "On the bursty evolution of blogspace," in *Proc. WWW*, 2003.

[19] J. Henning, "The blogging iceberg," October 2003.
http://www.perseus.com/blogsurvey/.

[20] E. Cohen, "Size-estimation framework with applications to transitive closure and reachability," *J. Computer and System Sciences*, vol. 55, pp. 441–453, 1997.

[21] A. Broder, S. Glassman, M. Manasse, and G. Zweig, "Syntactic clustering of the web," in *Proceedings of the 6th World Wide Web Conference*, 1997.

[22] B. Krishnamurthy and J. Wang, "On Network-aware Clustering of Web Clients," in *Proceedings of ACM Sigcomm*, August 2000. http://www.research.att.com/~bala/papers/sigcomm2k.ps.

[23] S. Hodgson, "The unpersons group blog," October 2003.
http://unpersons.net/archives/000055.html.

[24] B. Krishnamurthy and M. Arlitt, "PRO-COW: Protocol Compliance on the Web—A Longitudinal Study," in *Proc. USENIX Symposium on Internet Technologies and Systems*, Mar. 2001.
http://www.research.att.com/~bala/papers/usits01.ps.gz.

[25] D. Mosberger and T. Jin, "httperf—A Tool for Measuring Web Server Performance," in *Proc. Workshop on Internet Server Performance*, pp. 59–67, June 1998.
http://www.hpl.hp.com/personal/David_Mosberger/
httperf.

[26] “Web 100.” http://www.web100.com.

[27] “Blogger help.” http://help.blogger.com/bin/answer.py?answer=808&topic=12.

[28] “Bloglines.” http://www.bloglines.com/topblogs.

[29] “Userland news aggregator.” http://radio.userland.com/newsAggregator.

[30] “Blogdex.” http://blogdex.net/about.asp.

[31] “Daypop.” http://daypop.com.

[32] “Mom finds out about blog.” http://www.theonion.com/3944/news3.html.