On the Bursty Evolution of Blogspace

Ravi Kumar
ravi@almaden.ibm.com
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

Jasmine Novak
jnovak@us.ibm.com
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

Prabhakar Raghavan
pragh@verity.com
Verity Inc.
892 Ross Drive
Sunnyvale, CA 94089

Andrew Tomkins
tomkins@almaden.ibm.com
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120

ABSTRACT

We propose two new tools to address the evolution of hyperlinked corpora. First, we define *time graphs* to extend the traditional notion of an evolving directed graph, capturing link creation as a point phenomenon in time. Second, we develop definitions and algorithms for *time-dense community tracking*, to crystallize the notion of community evolution.

We develop these tools in the context of *Blogspace*, the space of weblogs (or *blogs*). Our study involves approximately 750K links among 25K blogs. We create a time graph on these blogs by an automatic analysis of their internal time stamps. We then study the evolution of connected component structure and microscopic community structure in this time graph.

We show that Blogspace underwent a transition behavior around the end of 2001, and has been rapidly expanding over the past year, not just in metrics of scale, but also in metrics of community structure and connectedness. This expansion shows no sign of abating, although measures of connectedness must plateau within two years.

By randomizing link destinations in Blogspace, but retaining sources and timestamps, we introduce a concept of *randomized Blogspace*. Herein, we observe similar evolution of a giant component, but no corresponding increase in community structure.

Having demonstrated the formation of micro-communities over time, we then turn to the ongoing activity within active communities. We extend recent work of Kleinberg [11] to discover dense periods of "bursty" intra-community link creation.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Measurement, Experimentation

WWW2003, May 20–24, 2003, Budapest, Hungary.
ACM 1-58113-680-3/03/0005.

1. INTRODUCTION

Weblogs (or *blogs* as they are more commonly known) constitute a fascinating artifact within the evolving web. Early hand-edited collections of blogs consisted of any page containing sequences of dated entries. Nowadays, most people think of blogs as pages with reverse chronological sequences of dated entries, usually containing a persistent sidebar containing profile information (and often other blogs read by the author) and usually maintained and published by one of the common variants of public-domain blog software. They tend to be quirky, highly personal, often consumed by regular repeat visitors and highly interwoven into a network of small but active micro-communities. In short, blogs are perhaps the most significant recent movement in end-user content creation on the web. We refer to the collection of blogs with their links as *Blogspace*.

There are at least two important reasons for the systematic study of Blogspace:

1. *Sociological reasons:* Blogspace differs from traditional web pages structurally because blogs represent concatenations of messages, as within newsgroups and bulletin boards, but authored by a single individual. However the more significant differences are more than structural: the culture of Blogspace focuses heavily on local community interactions between a small number (say, between 3 and 20) of bloggers. Members of the informal community might list one another's blogs in a "blogroll" and might read, link to, and respond to content on other community members' blogs. Often, these sequences of responses take place during a brief burst of heavy activity as an interesting topic arises, jumps to prominence, and then recedes. Naturally, this leads to the question: can we experimentally observe and model this highly dynamic, temporal community structure?

2. *Technical reasons:* Traditional studies of the web and the web graph make use of a static snapshot derived from a crawl. All such work raises the natural question: what happens over time? A number of works

have begun to address this question through creation and analysis of a series of snapshots of the data [4, 15, 8, 3]. The development of tools and methods to analyze these snapshots is therefore a timely endeavor. However, Blogspace offers an additional technical advantage over such approaches—if data is recrawled with a certain frequency, there is no notion of the precise point in time a page or link was created/updated. In contrast, Blogspace offers us a ready-made view of evolution in *continuous time:* as each blog adds an entry (together with links), there is a time stamp associated with that event. By automatically extracting these time stamps we can piece together a view of Blogspace evolving continuously from the beginning of blog archiving to the present. We should stress that time is absolute (not merely relative as in a sequence of crawls). Our work focuses on connectivity evolution and on temporally concentrated *bursts* (in the sense of Kleinberg [11]) in this evolution of Blogspace.

1.1 Main contributions

1. We introduce a combinatorial object we call a *time graph* (Section 3.1) for the study of graphs that evolve in continuous time. We build the *blog graph*—the time graph for Blogspace—by automatically extracting dates from blog page entries (Section 4).

2. We define a notion of communities in Blogspace and extend Kleinberg's notion of temporal bursts in a sequence of documents [11] to sets of blogs and the links between them, developing a notion of bursty communities of blogs that are topically *and* temporally focused (Section 3).

3. We conduct a series of experiments that develop properties and views of the graph induced by Blogspace as a function of time, showing the development of macroscopic and microscopic community structure, and the evolution of burstiness (Section 5).

4. We show that Blogspace underwent a transition behavior around the end of 2001, and has been rapidly expanding over the past year, not just in metrics of scale, but also in metrics of community structure and connectedness. This expansion shows no sign of abating, although measures of connectedness must plateau within two years (Section 5).

2. BACKGROUND

In this section we first provide some background material on graphs, communities, and burst analysis of events (Section 2.1). Next we review the world of blogs and argue that blog communities and web communities are different (Section 2.2).

2.1 Preliminaries

2.1.1 Graphs

A *directed graph* $G = (V, E)$ consists of a set V of *nodes* and a set E of *edges*, where each edge is an ordered pair of nodes. The *in-degree* of a node u is the number of nodes v such that $(v, u) \in E$; the *out-degree* of u is the number of nodes v such that $(u, v) \in E$. There is a *directed path* from u to v in G if there is a sequence of nodes $u = w_1, \ldots, w_k = v$ such that $(w_i, w_{i+1}) \in E$ for $1 \leq i < k$. A *strongly connected component* (SCC) of G is a subset of nodes such that for any ordered pair of nodes in the subset, there is a directed path from the former to the latter. An *undirected graph* $G = (V, E)$ consists of a set V of nodes and a set E of edges, where each edge is an unordered pair of nodes.

2.1.2 Communities

The notion of communities in the web graph (called *web communities*) was defined in [13] and the problem of extracting web communities was studied in [13, 12, 9]. Kumar et al. [13] detected communities by enumerating all bipartite cliques (up to a certain size) in the web graph. This approach was motivated by the co-citation phenomenon rampant on the web [10]. Their hypothesis was that any topically focused community on the web is likely to contain a dense bipartite subgraph (the *signature*) and almost every occurrence of the signature corresponds to a web community. Flake et al. [9] adopted a more sophisticated definition of a web community based on network flow. Section 3.2 describes our approach to the community extraction problem for Blogspace.

2.1.3 Bursts

We provide a brief review of Kleinberg's recent work on identifying bursts in a stream of events [11].

An event might correspond (for instance) to the appearance of an email containing particular keywords (such as "NSF grant"—a running example in [11]). The crucial step is to model such bursts so that they can be identified efficiently. Care must be taken to avoid identifying a large number of short spurious bursts or fragmenting long bursts into many smaller bursts.

Kleinberg's approach is to model the generation of events by an automaton that is in one of two states, "low" and "high." The time gaps between consecutive events are distributed independently according to an exponential distribution whose parameter depends on the current state. Thus the high state is hypothesized as generating bursts of events. There is a cost associated with any state transition to discourage short bursts. Given an event stream we seek to find a low cost state sequence that is likely to generate that stream. Finding an optimal solution to this problem can be accomplished by dynamic programming.

One final extension is required. Consider the case in which each event in a sequence is either *relevant* or *irrelevant*. Kleinberg extends his basic two-state model to this case as well. This augmented model generates events with a particular mix of relevant and irrelevant events according to a binomial distribution. A sequence of events is considered bursty if the fraction of relevant events alternates between periods in which it is large and long periods in which is small. Kleinberg defines a measure of weight associated with each such burst and solves the problem of enumerating all the bursts by order of weight.

2.2 Blogs

According to slashdot (`http://slashdot.org/features/99/05/13/1832251.shtml`), blogs are "... a new, personal, and determinedly non-hostile evolution of the electric community. They are also the freshest example of how people use the Net to make their own, radically different new me-

dia." Historically blogs date back to 1996, but they exploded into popularity during 1999 with the emergence of blogger (`http://www.blogger.com`) and other easy-to-use publishing tools. During 2000, they caught the public eye and articles began to appear in e-zines and forward-looking publications. Most recently in 2002, a Newsweek article (`http://stacks.msnbc.com/news/795156.asp`) appeared estimating the number of weblogs to be half a million, and discussing the emerging culture of blogspace:

> While weblogs had always included a mix of links, commentary, and personal notes, in the post-Blogger explosion increasing numbers of weblogs eschewed this focus on the web-at-large in favor of a sort of short-form journal. These blogs, often updated several times a day, were instead a record of the blogger's thoughts: something noticed on the way to work, notes about the weekend, a quick reflection on some subject or another. Links took the reader to the site of another blogger with whom the first was having a public conversation or had met the previous evening, or to the site of a band he had seen the night before. Full-blown conversations were carried on between three or five blogs, each referencing the other in their agreement or rebuttal of the other's positions.

It is precisely this type of intense interaction that is at the core of Blogspace[1] and that we wish to analyze algorithmically.

2.2.1 Bursty communities of blogs

At first blush, blog communities appear similar to web communities studied in earlier work [13, 9]. But there is a distinctly different flavor to blog communities, both qualitatively and (as we develop in subsequent sections) quantitatively: these communities exhibit striking temporal characteristics. Within a community of interacting bloggers, a given topic may become the subject of intense debate for a period of time, then fade away. These bursts of activity are typified by heightened hyperlinking amongst the blogs involved—*within a time interval*. Thus it no longer suffices (as in [13, 9]) to extract subgraphs that are signatures of communities; rather, we must extract such signatures while simultaneously identifying a time interval within which this hyperlinking is concentrated. Note that a subgraph indicative of a community of interest (in the traditional sense) may exist amongst a set of blogs, without ever achieving this temporal focus. Conversely, heavy linkage within a short period may appear less significant when viewed over a long time span—suggesting that the criterion for inferring that a pattern of links is a community be less stringent than for a static graph.

Identifying such temporal bursts is inspired by Kleinberg's recent work that was outlined in Section *2.1.3*. In Section 3 we extend Kleinberg's work to envelop *sets* of blogs inducing a bursty community through temporal bursts of hyperlinking. While deferring this formal development and experimentation to Section 3, we now give two examples of the bursty phenomena unearthed by our algorithms.

> In a community of blog poets, a burst occurs when one member Firda (`http://www.wannabegirl.org`) posts a series of daily poems about other bloggers in the community and includes links to their blogs. This burst occurs from March-April of 2000—for example `http://www.wannabegirl.org/2000_03_01_log-archive.php` from March 2000 contains poems and links to `http://trenchant.org/webloglog`, `http://www.premiumpolar.com/blog`, and `http://www.swallowingtacks.com`, all members of the community.

> In another community, a blogger Dawn (`http://up_yours.blogspot.com`) hosts a poll to determine the funniest and sexiest blogger. She conducts interviews with other bloggers in the community, of course listing their sites (see `http://up_yours.blogspot.com/2002_05_19_up_yours_archive.html`). She then becomes obsessed with one of the other bloggers Jim, which spurs comments by many others in the community (see `http://jimspot.blogspot.com/2002_07_28_jimspot_archive.html`)[2].

3. APPROACH

In this section we first define the notion of time graphs which will be be the basis for studying Blogspace. Time graphs can also be used to study different evolving graphs such as the web, e-mail graphs, call graphs, newsgroup graphs and so on. We anticipate that they will prove to be of considerable independent interest for many other mathematical and algorithmic studies.

Next we focus on tracking bursty communities on the time graph induced by blogs, henceforth the *blog graph*. We accomplish this by adopting a two-step approach:

1. Community extraction (Section 3.2): We extract dense subgraphs from the blog graph; these correspond to all potential communities (whether or not bursty).

2. Burst analysis (Section 3.3): Building on the work of Kleinberg [11] on bursts in event streams, we perform a burst analysis of each subgraph obtained in step 1 to identify and rank bursts in these communities.

The reason for this two-step approach is that the problem we wish to solve is somewhat harder than that addressed by Kleinberg. Whereas the elementary events he considers have simple, local characterizations (e.g., does an email contain a given keyword?), our setting does not afford such locality. A bursty community is not characterizable in terms of a single blog or edge in the time graph. Rather, it entails an analysis of the entire blog graph. Ideally, we must simultaneously identify subsets of blogs as communities together with bursts in the events relevant to this subset. We instead break this down into our two-step sequence; avoiding this two-step process remains a challenging open problem.

[1]For example, there is a "blogathon" held by `http://www.blogathon.org` once a year in which people blog for 24 hours straight for charity. Sponsors donate money and then during the blogathon, bloggers update their blogs every 30 minutes for an entire day.

[2]Dawn's blog and those of her community may not be suitable for all ages.

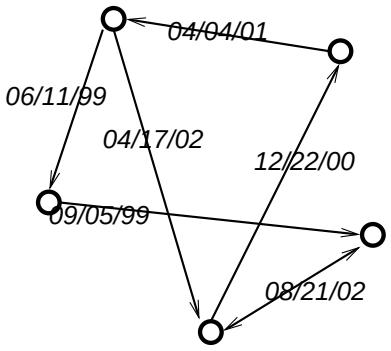

Figure 1: A typical signature of a blog community.

3.1 Time graphs

We now introduce what appears to be a novel combinatorial object: the time graph. A *time graph* $G = (V, E)$ consists of

1. A set V of nodes where each node $v \in V$ has an associated interval $D(v)$ on the time axis (called the *duration* of v)
2. A set E of edges where each $e \in E$ is a triple (u, v, t) where u and v are nodes in V and t is a point in time in the interval $D(u) \cap D(v)$.

A node v is said to be *alive* at time t if $t \in D(v)$. The interpretation is that each edge is created at a point in time at which its two end-points are alive. The definition naturally allows for directed time graphs. Note that in contrast to the well-established algorithmic study of dynamic graphs (see, for instance, [5]), edge events have real-valued time stamps.

Let $G = (V, E)$ be a time graph. The *prefix* of G at time t is also a time graph $G_t = (V_t, E_t)$ where $V_t = \{v \in V \mid D(v) \cap [0, t] \neq \emptyset\}$. Likewise, $E_t = \{(u, v, t') \in E \mid t' \leq t\}$.

3.2 Algorithms for community extraction

In the context of time graphs and blogs in particular, we adopt a more relaxed definition of communities than in [13]. There are at least two motivations for doing this:

1. Compared to the web, blogs are not characterized by the strong distinction between "authority-type" and "hub-type" pages [10]. Every node in the blog graph corresponds to a 'human being'; this is in contrast with the web where pages can be loosely classified as 'people' (the hubs) and 'topics' (the authorities).
2. In contrast with the web-scale experiments of [13, 12], the scale of our work here permits us to operate entirely in memory without streaming the data from disk. As a result, it is feasible for us to seek dense (rather than complete) subgraphs as community signatures.

We therefore consider the undirected version of the blog graph and say that a dense subgraph is a signature of a blog community. We will make the notion of a dense subgraph more precise later (see Figure 1 for a simple example).

Unfortunately, finding the densest subgraph in an undirected graph is NP-hard and appears notoriously difficult to even solve approximately [7]. We therefore resort to heuristics that are simple, efficient, and effective. The blog graphs we deal with are small enough that we can perform all the operations in memory, in contrast to the earlier work of Kumar et al. [13].

Preprocessing. First, following [13], we adopt the notion that pages linked-to by an enormous number of other pages are too well-known for the type of communities we seek to discover; so, we summarily remove all pages that contain more than a certain number of in-links. Next, we remove templates from the graph—this has implications for the extraction of communities and also for the burst analysis described below. The details of template identification and removal specific to blog graphs are presented in Section 5.

Our algorithm consists of two steps—pruning and expansion. Pruning corresponds to identifying the seed of a community and expansion corresponds to growing the seed to a dense subgraph that forms the signature of a community. We adopt the convention that that a node can participate in at most one community.

Pruning. We adopt the following algorithm for pruning, based roughly on the original work of Agrawal et al. [1] and the approach of Kumar et al. [13]. The graph is first scanned for all vertices of degree at most two. Vertices of degree zero and one are removed, and vertices of degree two are checked to determine whether they participate in a K_3; that is, whether their two neighbors are connected. If so, they are passed through as a seed to the expansion step (described below) and the resulting community is output and removed from the graph, if it passes a certain threshold.

After the entire graph has been processed in this manner, certain vertices that previously had degree three or more will now have degree two or less; hence, the pruning step is repeated several times (specifically, three times in our case).

Following the pruning passes the graph is processed greedily as follows. An arbitrary edge of the graph is extracted and then grown into a community according to the expansion algorithm given below. If the resulting community passes a size threshold, then it is output. In either case, it is removed from the graph and the process is repeated until there are no remaining edges.

Deletion of vertices is performed by appending the vertex to a "delete list," and checking this list whenever edges are extracted from the edge data structure. Once the delete list becomes sufficiently large, it is "garbage collected" back into the graph.

Expansion. The aim of the expansion step is to grow the seed into a set of nodes that constitute a potential community. First, it determines the vertex that contains most links to the current community. If that vertex contains at least t_k such links where t_k is a threshold depending only on the size k of the community grown so far, then it is added to the current community and the process repeats.

3.3 Burst analysis

In our context, the goal is to identify communities that are bursty in the blog graph. There is a natural interpretation of arrivals of edges in the blog graph as an event stream. Recall from Section 3.2 that a community corresponds to an undirected dense subgraph. Given a specific community

$C = (V_C, E_C)$, the relevant events correspond to the arrivals of the edges in E_C. Then, applying Kleinberg's algorithm [11] (see Section *2.1.3*), we can obtain the weight of every burst of C. We apply this algorithm for each extracted community in the graph.

4. METHODOLOGY

4.1 Data acquisition

We collected the data for the blog graph from the following seven popular blog sites: `http://www.blogger.com`, `http://www.memepool.com`, `http://www.globeofblogs.com`, `http://www.metafilter.com`, `http://blogs.salon.com`, `http://www.blogtree.com`, and `Web_Logs` subtree of Yahoo!

Some of the above sites list members explicitly in a directory while others categorize members by articles or topics. We crawled these sites to obtain a list consisting of 24,109 urls corresponding to blog member homepages. For each of the blog members, we crawled both their homepages and their archives; while the homepages represent the latest entries in the blog, the archives contain historical entries. Thus, for each blog, we are able to extract the entire detailed history of every link ever added to the blog, with the exact time at which it was added.

We used a very simple heuristic to identify if an out-link from the blog member's homepage was indeed an archive link: the url must contain the prefix 'archiv' or some indication of date and must have an indication of the blog member (name, id number, etc.). The out-links of a blog member are identified to be the (multiset) union of the out-links from the homepage of the blog member and each of the archive page. The *blog graph* is now defined to contain nodes that correspond to blog members and a link from node p to node q if blog member p created a link to blog member q at some point in time. The resulting graph consisted of 22,299 nodes, 70,472 unique edges, and 777,653 edges counting multiplicity. Observe that the average edge multiplicity of Blogspace is 11, a reflection of the highly interactive nature of linking. Note that we have not associated time information with each edge as yet; this is done in the next step.

4.2 Time graph generation

The subject of extracting specified entities from documents is a subject of on-going research (see, for instance, [14]). In our case, the entities correspond to valid dates. Because we were focused on date extraction from blogs in particular, we adopted a blog-specific scheme. Most blogs are published using a blog publishing software package (say, `blogger` that is available from `http://www.blogger.com`), and therefore specify dates in a uniform format. However, there will also be additional dates that occur textually within the blog, and we must be careful not to be misled into believing that these dates represent a new journal entry.

We created a broad set of date patterns based on various numeric and alphanumeric date formats, including only partially-specified dates (i.e., "September 1", which does not contain a year). We applied these patterns to the text of each blog page, and for each spotted occurrence we noted the family of pattern that matched, and the surrounding context of the reference. We then processed the entire sequence of extracted dates to determine whether a particular pattern family and/or context was repeated frequently enough. If so, we adopted the dominant scheme as the dates inserted by the blog publication software. We introduced some special logic to match contexts for well-known blog publishing tools such as `blogger`. Finally, we back-filled missing year information into partially-specified dates to complete them.

We did not implement the additional heuristic of checking that all dates believed to be journal entries form a decreasing or possibly increasing sequence, but this heuristic might have increased our confidence. Using the existing algorithm, we were able to associate dates with about 90% of the links extracted from post-template-removal blog pages.

The remaining 10% of edges occur due to various template links and "blogrolls"—a list of links to fellow bloggers' sites. We assigned a time of 0 for these links. We consider a blog to be alive since the earliest non-zero time tag of its out-links.

Bucketing the time in terms of the number of months since January 1999 lets us construct a sequence of 47 prefixes of the blog graph. We refer to these 47 graphs as *prefix graphs*.

4.3 Tuning the algorithms

We now discuss the specific settings of various parameters while running our algorithms on the blog graph.

Detecting templates in web pages is in itself an active area of research (see, for instance, [2]). It is particularly acute in the case of blogs for the following reason. Bloggers often modify a profile that appears as a template on all archive pages corresponding to the blog, and may include a series of links[3]. We wish to avoid our algorithm being misled into thinking that the archive corresponding to a particular date includes those links, when in fact the archive may be several years old and the template may have been created only yesterday.

We adopt the following simple heuristic for removing templates: remove any sequence of three or more links occurring in a blog for three or more days. To be conservative, we also removed links with time 0. As in [13], we also removed any node with in-degree more than 1000 from consideration in the community identification.

In the expansion step of the community extraction, the thresholds t_k were determined heuristically as follows: edges must grow to triangles; communities of size up to six will only grow vertices that link to all but one vertex; communities of size up to nine will only grow vertices that link to all but two vertices; communities up to size 20 will grow only vertices that link to 70% of the community; and larger communities will grow only vertices that link to at least 60% of the community.[4]

In burst analysis, we identified the links in each community as relevant events (as in Section *2.1.3*) and divided them into chronological batches according to the week each link was added. For each community identified in the previous step, we calculated the number of links created between

[3]Recently, the Rich Site Summary (RSS) XML format (see `http://www.webreference.com/authoring/languages/xml/rss/intro`) has become a popular means to announce "What's new" in the blogging community.

[4]It is possible that many vertices have at least t_k links to the current community. The algorithm as specified will add only the best such vertex, but for much larger datasets, it may be necessary to expand the current community by more than one vertex at a time. In such cases, we recommend no more than doubling the size of the current community at each step, to avoid adding large numbers of disjoint pages linked to a small central core.

blog members in the community during each week and the total number of links between pages in the community, to use as input to a two-state automaton. Each state of our automaton corresponds to a different fraction of relevant link generation: a lower rate during calm periods and a higher rate during bursty periods. By adjusting the scaling parameter which determines how much the high rate differs from the low rate, we were able to control the length of typical bursts. We experimented with various values for this parameter, and chose a value which resulted in the majority of bursts ranging from one week to several months.

5. RESULTS

We begin in Section 5.1 with an analysis of the evolution of structural properties of the time graph, as shown by analysis of the sequence of prefix graphs (as defined in Section 4). This analysis shows surprising behavior: around the end of 2001, the macroscopic structure (as measured by the formation of a giant component) and the microscopic structure (as measured by the formation of a large number of small communities) of the graph began to change dramatically. In Section 5.2 we argue that the change cannot be explained simply through the size, density, and link arrival pattern of the graph, but in fact arises only because of the particular linking decisions made by bloggers. In light of this observation Section 5.3 then presents our analysis of bursty behavior within the blog communities we have extracted, and shows that this burstiness is a fundamental property of link creation in blogspace.

5.1 Analysis of prefix graphs

Degree distributions. We first study the degree distributions of prefixes of the blog graph. The results are shown in Figure 2. Each line in the figure represents a prefix graph of the full time graph corresponding to a snapshot of the graph at a particular point in time. Higher lines correspond to more recent snapshots. The upper graph gives the in-degree distribution; that is, the number of pages in the time graph with a given in-degree. The lower graph gives the out-degree distribution. As the figure illustrates, the distributions remains fairly stable over time, increasing uniformly in y value due to the growth in size of the graph, but retaining the same overall shape. The later curves also become smoother, and it is possible to note that the tail of the curves (to the right of the graph, corresponding to nodes of higher degree) tracks fairly well to the power law curve with exponent -2.1.

Connectivity. We also study the evolution of the strongly connected component (SCC) in the prefix graphs. The results are shown in Figure 3. For each of the three largest strongly connected components, the figure shows what *fraction* of the nodes in the prefix graph are part of that SCC at each point in time. The results here are quite dramatic. In January of 1999, at the beginning of our study, the number of blog pages is significant but there is no strongly connected component of more than a few nodes. Around the beginning of 2000, a few components representing 1-2% of the nodes in the graph appear, and maintain roughly the same relative size for the next year. But up to this point, blogspace is not a coherent entity—the overall size has grown but the interconnectedness is not significant. At the start of 2001, the largest component begins to grow in size relative to the rest of the graph, and by the end of 2001 it contains about 3%

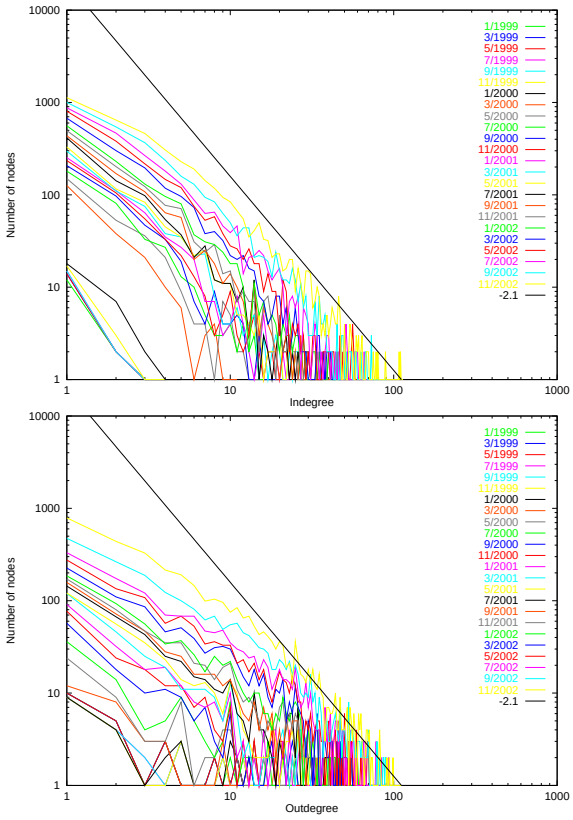

Figure 2: Evolving degree distributions

of all nodes. In 2002, however, a threshold behavior arises, and the size of the component increases dramatically, to over 20% by the present day. This giant component still appears to be expanding rapidly, doubling in size approximately every three months. Clearly this growth cannot continue and must plateau within two years.

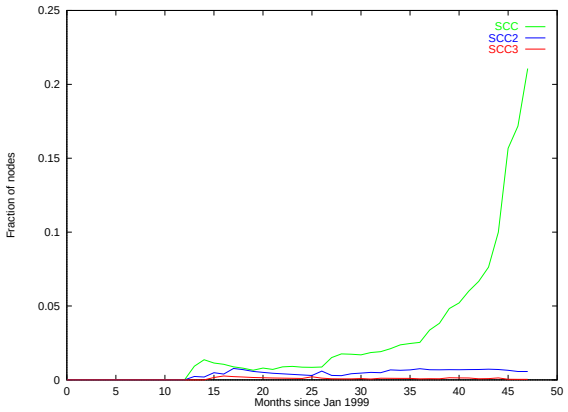

Figure 3: SCC evolution

Communities. We now turn to an analysis of how many pages take part in a "community," according to the definition implied by the community extraction algorithm of Section 3.2.

Table 1 shows the number of communities of each size discovered by the algorithm on the underlying graph. Figure 4

Size	3	4	5	6	7	8	9
#	143	165	79	14	2	1	5

Table 1: Distribution of community sizes in the blog graph

shows the results of applying the same algorithm to the prefix graphs. The upper figure plots for each time interval the total number of communities in the prefix graph, and the total number of nodes that participate in one of those communities. The lower figure plots for each time interval the fraction of nodes that belong to some community. If this fraction is large, one can view the space of all blogs at that time as consisting as a set of small inter-networking communities, rather than a set of standalone pages.

These graph show the same striking pattern as earlier graphs in this section: a period of minimal structure during 1999 and 2000, slow growth in 2001, and then rapid growth in 2002. To conclude, the degree distributions match ear-

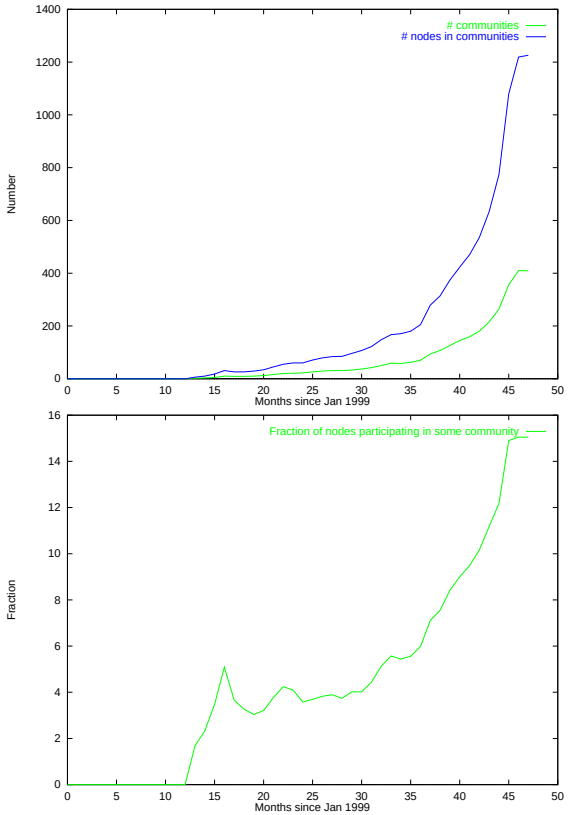

Figure 4: Community evolution

lier observations from many communities, and do not represent a surprise. The analysis of the largest SCC (a macroscopic phenomenon) and of communities (a microscopic phenomenon) does represent a surprise: by both measures, a fundamental change occurred in blogspace approximately one year ago, and we are still experiencing the results of that transition.

To assess whether this observed behavior does in fact stem from the sociology of blogspace, we must first study the alternate possibility: namely, that the emergence of a giant

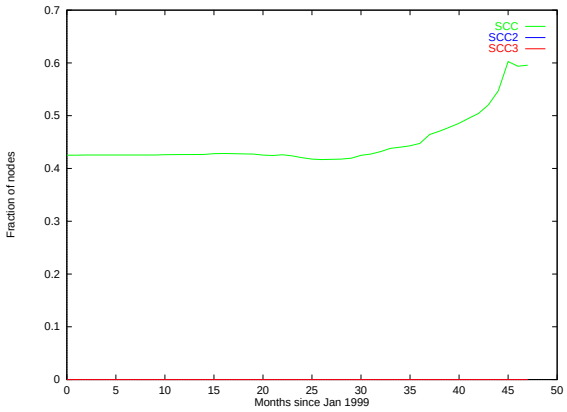

Figure 5: SCC evolution in randomized Blogspace

component and a strong community structure would occur naturally when the graph reached a certain size and density. We now address this question.

5.2 How random is the blog graph?

We wish to determine whether the prefix graph behavior we have seen is caused by a phenomenon similar to that of a time-evolving version of Erdös-Rényi random graphs [6], but tailored to produce graphs that match Blogspace in the source and arrival time of every edge. To study this, we create a graph derived from Blogspace called *randomized Blogspace*. This graph is identical to the blog time graph, except that the destination of every edge is replaced by a uniformly chosen random node. Thus, the arrival time of each edge, the number of edges at each time, and the exact profile of when a page chooses to add a link is left untouched. The only difference is the destination of each new link.

To be precise, we sort the edges of the blog graph according to time (ties are broken arbitrarily). We scan the edges sequentially and change each destination to be a node uniformly chosen from among the nodes that have already been seen. Note that this preserves the times at which links arrives at all sources, and modifies only the destinations of those links.

Figures 5 and 6 plot the same quantities as Figure 3 and Figure 4 for randomized Blogspace instead of the original blog graph. Since we included the time 0 edges, there is a substantial SCC to begin with. As time progresses, this SCC seems to have a threshold growth as in blog graph case (this is how a random graph would behave as well). For completeness, we also evaluated the growth of the giant SCC without the time 0 edges; initially it was of course much smaller, but it exhibited a similar threshold behavior and became a larger fraction of the overall graph during the last timestep than in Blogspace.

However, comparing to Figure 4, we see that the number of nodes in communities for randomized Blogspace is an order of magnitude smaller than for Blogspace, indicating that the community formation in Blogspace is not simply an emergent property of the growth of the graph. On the other hand, comparing Figure 3 to Figure 5 shows that the SCC in randomized Blogspace grows much faster than in the original blog graph.

So randomized Blogspace actually attains a large strongly connected component faster than Blogspace does; however,

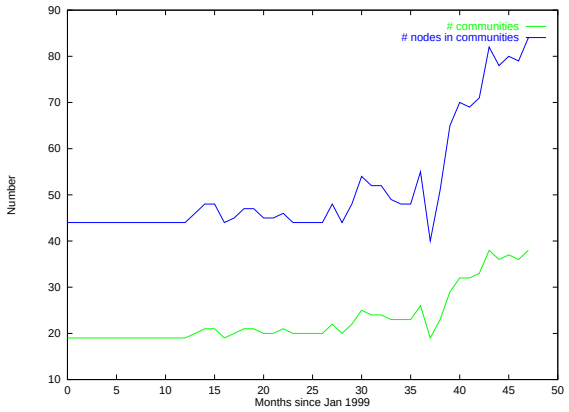

Figure 6: Community evolution in randomized Blogspace

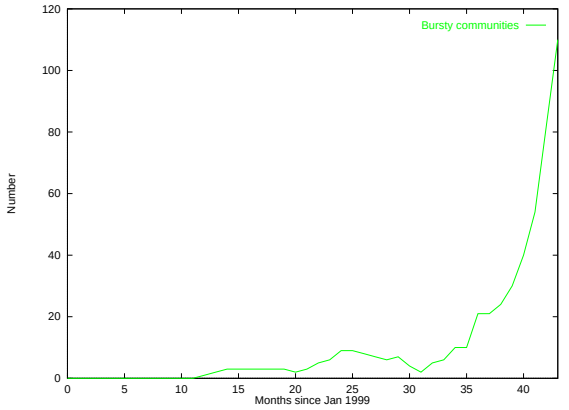

Figure 7: Burstiness of communities

it does not attain significant community structure. If bloggers added links to other blogs without reference to topicality, the graph would still become well-connected, but it would not exhibit the striking community focus that characterizes Blogspace.

5.3 Burstiness in blog communities

Figure 7 shows the number of communities that are in the high state during each time interval, as described in Section *2.1.3*. The x axis again shows time, but stops several months before our most recently crawled pages, because we cannot effectively evaluate the number of bursts occurring during the present without context about the near future. Again, consistent with our earlier observations, there is a spurt of bursty activity toward the end of 2001 that continues through to the present.

Interestingly, the increase in number of bursts is not explained by the increase in number of communities alone. Not only have the number of communities in Blogspace been growing over the last year, the average burstiness of each individual community has also been growing. This suggests that the transition behavior we have observed in all our temporal analyses is in fact correlated with a change in the behavior of the bloggers themselves. For whatever reason, perhaps because of the richer set of available communities, or the change in the population of Blogspace, content creators have increased their participation in bursty community activity over the last year, and the trend shows no sign of slowing.

5.4 Anecdotal evidence for the effectiveness of community and burst extraction

We explored a large number of communities and of bursts within communities using a web-based tool we created for the purpose. The communities found by the extraction algorithm are almost universally on topic. In all cases we examined, the communities contained many internal links, and it was usually clear what bound the members together: be it an interest in Flash, the law, or library science.

Periods of bursty behavior require a deeper investigation into blog content. In some cases, a burst occurs due to a spate of activity by one or two bloggers during the time period, as when during August and September of 2002 blogger Karen (`http://www.tempestinateapot.org`) started linking to her sister's blog (`http://www.ericalucci.com`) several times a day. Other bursts are the result of many members of the community contributing new links to each other. Although we are not always able to determine the reason for the intense period of linkage, in most cases there is a clear identifiable event or set of events. As the following closing example shows, the amount of information in a blog burst, and the window it gives into the lives of the bloggers, can be startling:

Alicia (`http://www.aliciadawn.com/blog`) is part of a group of artists in Seattle who form a blogger community. She's involved in fringe theater, and some of the other members are in a band together (see June 28, 2002 on `http://www.articulatebabble.org/archives/2002_06.php`). Several events surround the bursty link behavior during the four months from June-Oct in 2002. Alicia decides to connect with old high school friends (see June 24, 2002 on `http://www.aliciadawn.com/blog/archives/2002_06.html`). She asks two members of the community to set up blogs for them (see June 10, 2002 on `http://www.aliciadawn.com/blog/archives/2002_06.html`), which they do (see July 13, 2002 on `http://melody.asc-soft.com/~enigma/blog/archives/2002_07.html`). The event generates a mini-burst of blogging. She then convinces two high school friends to visit Seattle on two different weekends. Lots of blogging covers what to show them when they visit (see August 5, 2002 on `http://www.aliciadawn.com/blog/archives/2002_08.html`), waiting at the airport (see July 14, 2002 on `http://melody.asc-soft.com/~enigma/blog/archives/2002_07.html`), picking them up at the airport, their reaction to Alicia's theater performance, and so on. A third event during this same period occurs when two members of the community get engaged. There's discussion about the engagement, and the beautiful kids they'll have (see June 28, 2002 on `http://www.jetlin.com/blog/archives/2002_06.html` and `http://www.aliciadawn.com/blog/archives/2002_06.html`).

6. CONCLUSIONS

In analyzing the space of weblogs, we have presented a detailed picture of a web publishing phenomenon in the midst of explosive growth. Around the end of 2001, Blogspace began a dramatic increase in connectedness, and in local-scale community structure. Within those local communities, it also began to exhibit dramatic increases in the occurrence of bursty link creation behavior.

Blogspace represents a clean, detailed, and measurable instance of a hyperlinked corpus in evolution, and is thus an excellent testbed for exploring evolutionary analysis, in addition to being of significant interest in its own right. The tools we have developed are applicable to other evolving hyperlinked corpora, including sequences of snapshots of the web.

Acknowledgments

We thank Jon Kleinberg for providing the burst analysis code. We thank Amit Kumar for bringing to our attention the popularity of RSS XML among bloggers.

7. REFERENCES

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proc. 20th Intl. Conf. on Very Large Data Bases*, pages 487–499, 1994.

[2] Z. Bar-Yossef and S. Rajagopalan. Template detection via data mining and its applications. In *Proc. 11th Intl. World-Wide Web Conference*, pages 580–591, 2002.

[3] K. Bharat, B. Chang, M. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *IEEE International Conference on Data Mining*, 2001.

[4] B. E. Brewington and G. Cybenko. Keeping up with the changing web. *Computer*, 33(5):52–58, 2000.

[5] D. Eppstein, Z. Galil, and G. Italiano. Dynamic graph algorithms. In *CRC Handbook of Algorithms and Theory of Computation, Chapter 22*. CRC Press, 1997.

[6] P. Erdös and A. Rényi. On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intez. Kozl.*, 5:17–61, 1960.

[7] U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[8] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. Crawling towards light: A large scale study of the evolution of web pages. In *Proc. 1st Workshop on Algorithms for the Web*, 2002.

[9] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *Proc. 6th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 150–160, 2000.

[10] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 2000.

[11] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Extracting large scale knowledge bases from the web. In *Proc. 27th Intl. Conf. on Very Large Data Bases*, pages 639–650, 1999.

[13] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for cyber communities. *WWW8/Computer Networks*, 31(11-16):1481–1493, 1999.

[14] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.

[15] The Internet Archive. `http://www.archive.org`.