Finding Patterns in Blog Shapes and Blog Evolution

Mary McGlohon, Jure Leskovec, Christos Faloutsos
Matthew Hurst†, Natalie Glance†

January 2007
CMU-ML-07-100

Carnegie Mellon

Finding Patterns in Blog Shapes and Blog Evolution

Mary McGlohon, Jure Leskovec, Christos Faloutsos*
Matthew Hurst, Natalie Glance†

January 2007
CMU-ML-07-100

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

* School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
† Neilsen Buzzmetrics, Pittsburgh, PA, USA.

This material is based upon work supported by the National Science Foundation under Grants No. IIS-0209107, SENSOR-0329549, EF-0331657, IIS-0326322, IIS-0534205, and also by the Pennsylvania Infrastructure Technology Alliance (PITA). Additional funding was provided by a generous gift from Hewlett-Packard. Mary McGlohon was partially supported by a National Science Foundation Graduate Research Fellowship, and Jure Leskovec was partially supported by a Microsoft Research Graduate Fellowship.

Keywords: social network analysis, principal component analysis, bursty behavior

Abstract

Can we cluster blogs into types by considering their typical posting and linking behavior? How do blogs evolve over time? In this work we answer these questions, by providing several sets of blog and post features that can help distinguish between blogs. The first two sets of features focus on the topology of the cascades that the blogs are involved in, and the last set of features focuses on the temporal evolution, using chaotic and fractal ideas. We also propose to use PCA to reduce dimensionality, so that we can visualize the resulting clouds of points.

We run all our proposed tools on the ICWSM dataset. Our findings are that (a) topology features can help us distinguish blogs, like 'humor' versus 'conservative' blogs (b) the temporal activity of blogs is very non-uniform and bursty but (c) surprisingly often, it is *self-similar* and thus can be compactly characterized by the so-called *bias* factor (the '80' in a recursive 80-20 distribution).

Feature-set Name	Description	Feature size
CASCADETYPE	Clusters blogs based on the structure of the cascades (conversations) in which they participate.	$44,791$ blogs and $8,965$ cascade types (= features). For each cascade type, we record the frequency of such cascade type (and take the logarithm).
POSTFEATURES6	Clusters blogs based on features aggregated from their individual posts.	$44,791$ blogs made up of $6,666,188$ posts, with 6 descriptive features.
BLOGTIMEFRACTAL	Measures for burstiness of temporal activity posting and linking features.	6 descriptive features, "bias" factors for inLinks, number of posts, etc.

Table 1: The three tools used in this work for characterizing blogs

1 Introduction

The blogosphere is often viewed as a social network. It consists of a community of users that interact with each other, forming links, cliques, and sub-communities. Therefore, like other social networks, one might expect members of the blogosphere to assume certain functions in shaping the overall community. There may be some particularly prominent members who start major conversations; and there may be others who are more active in gathering content from many conversations.

Our goal is to find patterns of blog behavior, either looking at the topology of the cascades of a blog, or the temporal activity of a blog (number of posts over time, number of in-links etc). In this work we want to answer the question "*what properties are most indicative of a given blog*"? Our goal is to propose a set of characteristics that may serve as a "blog profile" for classification.

We propose three sets of features for classifying blogs: CASCADETYPE, POSTFEATURES6, and BLOGTIMEFRACTAL. CASCADETYPE has about $9,000$ features, trying to capture the "shapes" of the conversations in which the blog participates in. Typical shapes are "stars" and "chains". POSTFEATURES6 uses a set of characteristics of posts (like in-links, conversation mass), to capture the behavior of the corresponding blog. Finally, we use the BLOGTIMEFRACTAL set of features to characterize the temporal behavior, which, as we show, is bursty and self-similar. Table 1 lists our proposed feature sets.

The rest of the paper is organized as follows: Section 2 gives a literature survey. Sections 3,4,5 describe the three proposed sets of features, and our results on the real dataset. In section 6 we discuss our findings and a possible application in blog ranking. Section 7 gives the conclusions.

2 Related work

2.1 Blogs and communities

Most work on modeling link behavior in large-scale on-line data has been done in the blog domain [1, 2, 15]. The authors note that, while information propagates between blogs, examples of genuine cascading behavior appeared relatively rare. This may, however, be due in part to the Web-crawling and text analysis techniques used to infer relationships among posts [2, 11]. Our work here differs in a way that we concentrate solely on the propagation of links, and do not infer additional links from text of the post, which gives us more accurate information. Finally, work in [18] identified many patterns in blog linking patterns and proposed a model for reproducing cascading behavior.

There has also been much work on the community structure of the blogosphere. Work on information diffusion based on topics [11] showed that for some topics, their popularity remains constant in time ("chatter") while for other topics the popularity is more volatile ("spikes"). There is also work on finding what blogs are the most influential. Authors in [15] analyze community-level behavior as inferred from blog-rolls – permanent links between "friend" blogs. Analysis based on thresholding as well as alternative probabilistic models of node activation is considered in the context of finding the most influential nodes in a network [13], and for viral marketing [20]. Such analytical work posits a known network, and uses the model to find the most influential nodes.

The authors of [1] also showed that sub-communities may assume different characteristics: in particular, for blogs during the 2004 election the liberal community was far less connected than the conservative one. In a related social network, the Usenet, Fiore *et al.* assigned roles that different users played based on a survey, and were able to identify some common network characteristics of these different roles [8].

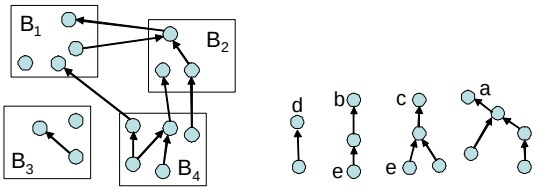
2.2 Information cascades

Information cascades, which will be described in detail in Section 3, are phenomena in which an action or idea becomes widely adopted due to the influence of others, typically, neighbors in some network [5, 9, 10]. Cascades on random graphs using a threshold model have been theoretically analyzed [26]. Empirical analysis of the topological patterns of cascades in the context of a large product recommendation network is in [19] and [17].

2.3 Burstiness and power laws

Extensive work has been published on patterns relating to human behavior, which often generates bursty traffic. Disk accesses, network traffic, web-server traffic all exhibit burstiness. Wang et al in [25] provide fast algorithms for modeling such burstiness. Burstiness is often related to self-similarity, which was studied in the context of World Wide Web traffic [6]. Vazquez et al [24] demonstrate the bursty behavior in web page visits and corresponding response times.

Self-similarity, fractals and *power laws* often appear together [22]. Power laws are laws of the form $y = x^a$, where a is the exponent of the power law. Probably the most famous such power law



(a) Blogosphere (b) Cascades

Figure 1: The model of the blogosphere *(*a*)*. Squares represent blogs and circles blog-posts. Each post belongs to a blog, and can contain hyper-links to other posts and resources on the web. From the blogosphere, we extract cascades *(*b*)*. Cascades represent the flow of information through nodes in the network. To extract a cascade we begin with an initiator with no out-links to other posts, then add nodes with edges linking to the initiator, and subsequently nodes that link to any other nodes in the cascade

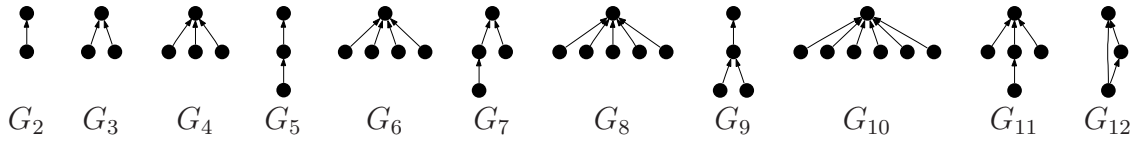


Figure 2: Common cascade shapes, ordered by the frequency in the dataset

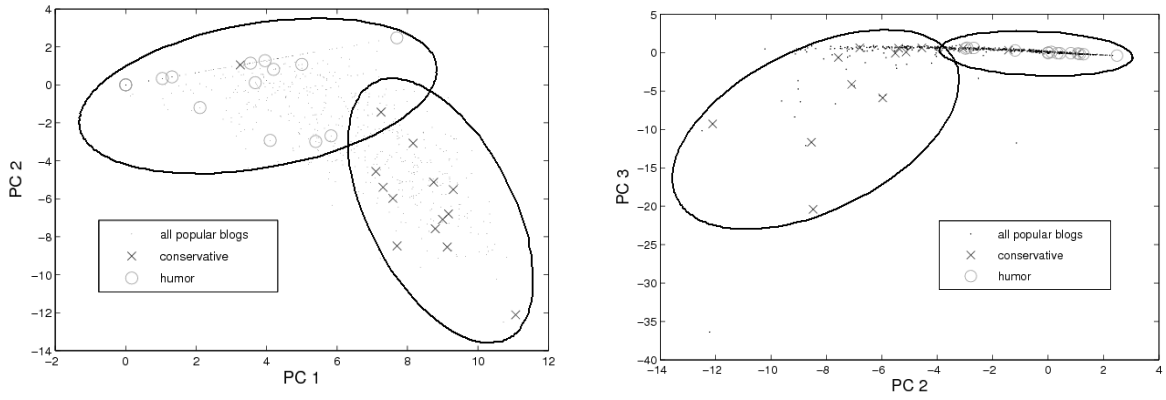
is the Zipf distribution [27]. Power laws on the topology of graphs have recently appeared, and specifically on the degree distribution of the web [4, 3, 16] and of the Internet [7].

3 Blog topology and roles

3.1 Preliminaries

The blogosphere is composed of blogs, which are further composed of posts. Posts then contain links to other posts and resources on the web. From the blogosphere, we extract information cascades, which are induced subgraphs by edges representing the flow of information. A cascade (also known as conversation tree) has a single starting post called the *cascade initiator* with no out-links to other posts (e.g. nodes a, b, c, d in Figure 1(b)). Posts then join the cascade by linking to the initiator, and subsequently new posts join by linking to members within the cascade. Figure 1(b) gives a list of cascades extracted from the network in Figure 1(a). Since a link points from the follow-up post to the existing (older) post, influence propagates following the reverse direction of the edges.

We also define a *non-trivial* cascade to be a cascade containing at least two posts, and therefore a *trivial cascade* is an isolated post. Figure 1(b) shows all non-trivial cascades in Figure 1(a), but not the two trivial cascades. Cascades form two main shapes, which we will refer to as *stars* and *chains*. A star occurs when a single center posts is linked by several other posts, but the links do not propagate further. This produces a wide, shallow tree. Conversely, a chain occurs when a root



(a) First vs. second PC

(b) Second vs. third PC

Figure 3: Principal components for blogs by CASCADETYPE labeled by topic. PC's were generated by analyzing a matrix of blogs by counts of cascade types. Note that there is a clear separation between conservative blogs *(*represented by red crosses*)*, and humorous blogs *(*represented with by circles*)*, both on axes of the first and second PC *(*a*)*, and on axes of the second and third PC *(*b*)*. Ovals indicate the main clusters

is linked by a single post, which in turn is linked by another post. This creates a deep tree that has little breadth. As we will later see most cascades are somewhere between these two extreme points. Occasionally separate cascades might be joined by a single post – for instance, a post may summarize a set of topics, or focus on a certain topic and provide links to different sources that are members of independent cascades. The post merging the cascades is called a *connector node*. Node e in Figure 1(b) is a connector node. It appears in two cascades by connecting cascades starting at nodes b and c.

Cascades are the basis of w hat we understand to be the patterns of diffusion of information through the blogosphere. They take on a number of different variants of chains and stars in shape. We enumerate these different shapes into types. A few of the more common types are shown in Figure 2. We propose to explore blogs based on the typical cascade shapes they take on.

We also define different characteristics of blogs. *In-link* and *out-link* represent links to and from a post or blog, and *depth* upwards or downwards represents the depth of the cascade tree.

Finally, we use *conversation mass*. Let T be the set of all cascades, B be the set of all bloggers, and P be the set of all posts. Let $T(b)$ be the subset of all conversations in which blogger b (in B) contributes at least one post.

Let $t \in T$ be an instance of a cascade. $t(p)$, for $t \in T$ and $p \in P$ is the subtree of the conversation t starting at post p. Define the conversation mass generated by post p as the number of posts in $t(p)$. Define the conversation mass for blogger B as the sum of the conversation mass of $t(p)$ over all t in $T(B)$, where p is the first post in t authored by blogger b.

In other words, the conversation mass for a blogger equals: the total number of posts in all conversation trees below the point in which the blogger contributed, summed over all conversation trees in which the blogger appears.

3.2 Principal component analysis

Given many vectors in D-dimensional space, how can visualize them, when the dimensionality D is high? This is exactly where Principal Component Analysis (PCA) helps. PCA will find the optimal 2-dimensional plane to project the data points, maintaining the pair-wise distances as best as possible. PCA is even more powerful than that: it can give us a sorted list of directions ("principal components") on which we can project. See [12] or [14] for more details.

3.3 Clustering blogs by CASCADETYPE

Our first experiments involved performing PCA on a large, sparse matrix where rows represented blogs and columns represented different types of cascades. Each entry was a count, and in order to reduce the variance, we took the log of each count. Our dataset consisted of $44,791$ blogs with $8,965$ cascade types.

It was of interest to impose social networks upon the blogs, based on what topics the blogs tended to focus on. We hand-classified a sample of the blogs in the ICWSM data by topic, and found that we could often separate communities based on this analysis. For the purposes of visualization we chose to focus on two of the larger communities, politically conservative blogs and "humorous" blogs (such as blogs for different web-comics and humorists). Figure 3(a) shows these blogs plotted on the first two principal components, and Figure 3(b) shows them plotted on the second and third principal components. Ovals are drawn around the main clusters. We notice a distinct separation between the conservative community and the humor community; this means that the two communities engage in very different conversation patterns.

3.4 Observations

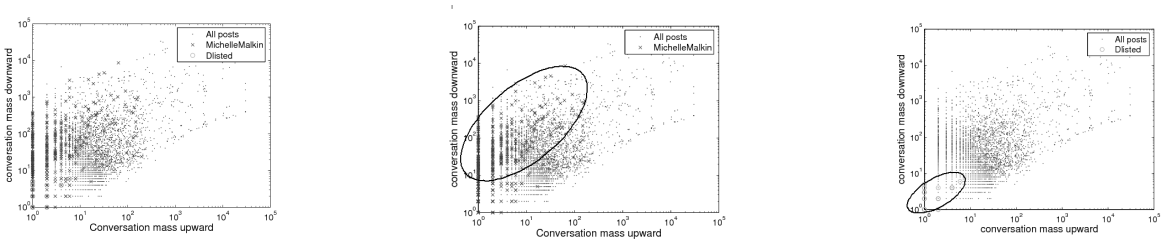
Based on our CASCADETYPE analysis, we make the following observations:

Observation 1 *Communities often cluster around the same types of cascades, with distinct conversation patterns.*

It seems that conservative blogs and the "humorous" blogs form separate clusters. We believe this is the case because conservative blogs tend to form deep, chainlike graphs whereas the humorous blogs form stars. Some similar observations may be made for other communities; we used these two because they were the most distinct. This result shows that blog communities tend to follow different linking patterns. We believe that by looking at a blog's cascade types that one can better make inferences about what community a blog might belong to.

Observation 2 *The number of trivial cascades that a blog participates in–that is, its number of solitary posts with no in- or out-links, may be a key indicator of its community.*

Removing the trivial cascades caused the clusters to become less clear, which indicates that these trivial cascades still play a significant role in the inferences one can make about that blog.



(a) All posts (b) MichelleMalkin (superimposed on all) (c) Dlisted (on all)

Figure 4: Conversation mass for posts, an aspect of POSTFEATURES6. The top figure shows the Dlisted and MichelleMalkin clusters superimposed over points for all posts. The next two show the clusters separately, superimposed on all blog points for reference. Ovals are drawn around the main clusters. Note that while there is overlap between posts features of two blogs, they have different centers. This tells us that different blogs maintain different means and variances in conversation masses

4 Post topology and roles

4.1 Clustering posts

We next sought to find how posts themselves behave. In order to do this, we performed PCA on a 6-column matrix. Each row represented a post, while the columns were as follows:

- Number of inlinks
- Number of outlinks
- Conversation mass upwards
- Conversation mass downwards
- Depth upwards
- Depth downwards

There were $6,666,188$ posts in the dataset. When we ran PCA, we found that the major two components that determined the blog's place in this space were conversation mass upwards and downwards. Therefore, we also plotted the posts on the two axes of conversation mass upwards and conversation mass downwards (See Figure 4. To illustrate, we have plotted all posts, with special markers for two distinct popular blogs, Dlisted [1] and MichelleMalkin [2]. We have circled the main clusters in the plots. Notice that while Dlisted and MichelleMalkin points overlap, their clusters are centered differently. The mean and variance of these clusters can serve as another viewpoint into the profile of a blog.

[1] `dlisted.blogspot.com`, a celebrity gossip blog.

[2] `www.MichelleMalkin.com`, a politically conservative blog.

4.2 Observations

Our POSTFEATURES6 analysis provided us the following observation:

Observation 3 *Posts within a blog tend to take on common network characteristics, which may serve as another means of classification.*

Individual posting patterns may serve as another way of clustering blogs, because different blogs maintain different posting patterns. We further discuss the use of two of the characteristics, in-link and conversation mass, in Section 6.

5 Time evolution and burstiness

Activity over time is bursty - see Figure 5(a-c). How do we quantify this burstiness, and how useful it is as a 'feature' to characterize the temporal behavior of a blog? In this section we propose to use tools from chaotic time series and fractals, exactly to answer the above question.

For clarity, we focus on $p(t)$ (posts over time) for the rest of this discussion. However, for our experiments, we also measure in-links $i(t)$, out-links $o(t)$, down-ward conversation mass $md(t)$, etc.

5.1 Definitions

From the few time plots shown in Figure 5(a-c), we see that real activity is far from uniform. Among the many measures for non-uniformity, we propose to use the entropy [23]. Recall that **entropy** on a random variable $\mathcal{X}$, (e.g., the outcome of a random dice) is defined as

$$H(\mathcal{X}) = -\sum_{i=0}^{N} p_i \log_2 p_i, \tag{1}$$

where p_i is the probability of each outcome ($1/6$ in the case of a dice) and N is the total number of possible outcomes (e.g. N=6, for the dice). $H()$ is close to 0 if the distribution is highly skewed while a uniform distribution gives the maximum value of $\log_2 N$ for H.

We propose to measure the non-uniformity of a time sequence like the number of posts $P(t)$ ($t = 1, \cdots, T$) as follows. Let $P_{total} = \sum_{t=1}^{T} P(t)$ be the total number of posts for the blog of interest, and let $p(t) = P(t)/P_{total}$ be the percentage of posts on day t. Then we use the entropy H_p of the time sequence $p(t)$ as a measure of non-uniformity:

$$H_p = -\sum_{t=1}^{T} p(t) \tag{2}$$

Thus, if the $p(t)$ activity is uniform over time, the value of its entropy H_p will be maximum. By looking at the (bursty) time-plots of Figure 5(a-f) we expect that the entropies will be much lower than the entropy maximum. It turns out that we have an even stronger way to characterize our

traffic, because it is *self-similar*: if we focus on a smaller sub-sequence, it will be statistically similar to the longer, mother-sequence it came from. Intuitively, if the original sequence has bursts and silences more bursts, so will the subsequence, with bursts-silences-bursts, at smaller scales.

5.2 The "b"-model: 80-20 recursively

How would such self similarity appear? it turns out that the recursive application of the 80-20 "law" results in such bursty *and* self-similar behavior. The "b"-model with *bias parameter* b generates activity as follows ($0.5 \leq b \leq 1.0$): If the total activity is, say, P total number of posts during the full interval of observation, and b=0.8 (80-20 law), the first half of the time interval receives b=80% fraction of the activity, and the second half receives the remaining 20%; the first quarter recursively receives 80% of the first half activity, and so on. That is, every sub-interval has exactly the same un-balance like its parent (and uncle, and grand-parent) intervals. Figure 6(a) illustrates the first few steps of the recursive generation of such bursty traffic. Figure 6(b) plots the generated traffic, with bias factor b=0.8, after 2^10 subdivisions. Notice how bursty the generated traffic is. Of course, we don't have to always favor the left sub-interval: we could occasionally flip our bias, to generate more natural-looking traffic.

5.3 Measuring the burstiness: the entropy plot

There are two questions: (a) How accurately does the *b-model* characterize our blog activities? and (b) How to measure the bias factor b, when we are given a real traffic (e.g., fraction of posts $p(t)$, per day).

The answer comes from the theory of fractals and disk traffic modeling, where the *entropy plot* [25] has been used successfully. Again, let's focus on the fraction of posts $p(t)$ per day. The idea is to compute the entropy H_p at the original resolution (1 day), as well as at coarser resolutions (sum of windows of size 2, 4, 8 days and so on). The way the entropy changes with the resolution answers both questions. We elaborate next.

For simplicity, suppose that the number of days T is a power of 2: $T = 2^r$. If not, we can zero-pad the sequence, or clip it to the highest power of 2. Let r stand for the original *resolution*, and let $H_p(r')$ denote the entropy at resolution r' ($0 \leq r' \leq r$). The sequence at resolution r is the original sequence, with duration $T = 2^r$; at resolution $r-1$, the sequence is the sum of successive, disjoint windows of size 2, with duration $T/2$. In general, at resolution $r-i$, we divide the original sequence into disjoint windows of size 2^i, sum them, and compute the entropy $H_p(r-i)$.

Clearly, for resolution 0, the whole sequence collapses to one number, '1', and its entropy is zero (the entropy of a completely unfair coin that always brings 'Heads').

Figure 6(c) gives an example. The horizontal axis is the *resolution* r' (0 for the whole interval, 1 for two halves, e.t.c.) and the vertical axis is the entropy $H_p(r')$ of the activity, as described above.

As discussed in [25], traffic generated by the b-model is self-similar, and its entropy plot is linear. Surprisingly, many of the blogs we examined showed activity that *also* resulted in linear entropy plots, in all features we tried: number of posts per day, number of in-links per day, etc., as shown in Figure 5 (g,h,i).

The linearity of the entropy plot gives evidence that a b-model may be a good model for our blog traffics, answering the first question we posed in this subsection. For the second question, how to estimate the bias parameter b, we have the following Lemma:

Lemma 1 *For traffic generated by a b-model, the slope s of the entropy plot, and the bias factor b obey the equation*

$$s = -b\log_2 b - (1-b)\log_2(1-b)$$

Proof: See [25]

Notice that bias b=0.5 corresponds to the uniform distribution (fifty-fifty splits for each sub-interval, and slope s=1 for the entropy plot) The higher the value of b, the more bursty is the time sequence. In the extreme case of b=1.0, all the activity is zero everywhere, except for a burst at one single day, and the slope s=0 for the entropy plot. The slope s is actually the so-called 'Information Fractal Dimension', which estimates the intrinsic dimensionality of a cloud of points (timestamps of posts, in our example): if the timestamps are uniformly distributed, the resulting cloud of timestamps has dimensionality s=1 (the whole line interval); if they are all on the same, single day, the cloud degenerates to point, with dimensionality s=0. Our real datasets have dimensionality between zero and one, like, e.g., the famous 'Cantor dust' dataset (remove the middle-third from the unit interval, and repeat recursively for the two pieces). Figure 7 shows the distribution of bias factors of the time-sequences.

5.4 Observations

Using the bursty view point and the "bias factor", we have the following observations:

Observation 4 *Most of the time series of interest are self-similar.*

Notice that they didn't have to be self-similar: the entropy plots could be parabolic, or piece-wise linear, or any other form than linear. Yet, most of them are indeed self-similar!

Observation 5 *Most of the bias factors are in the 70% range, that is, much more bursty than uniform (Poisson).*

The uniform distribution (that is "Poisson arrivals") would lead to bias factors around 50% (fifty-fifty splits). Our measurements are not even close to that. In retrospect, it makes sense, because blog activity is bursty: a few posts "hit a nerve" and attract a lot of interest, while the vast majority of posts does not.

6 Discussion

The methods chosen in this work were decided mainly for simplicity, as the main goal was to present ideas for some blog characterization. For CASCADETYPE and POSTFEATURES6 we ran PCA after taking the log counts. There are other methods available for reducing variance, however,

we chose log for the sake of simplicity. It may be of interest to use different forms of TF-IDF, a method often used in text mining. A description of TF-IDF is provided in [21].
We have analyzed many characteristics of blogs, based on conversation patterns, post features, and post patterns over time. From this basis, given a blog, we can infer a number of things about that blog based on these metrics.

6.1 Ranking blogs by in-link count vs. conversation mass

Another useful application of blog classification is ranking blogs for search results. The usual method of ranking is in-links. However, simply counting the number of in-links does not capture the amount of "buzz" a particular post or blog creates. We argue that conversation mass is another important feature that is not necessarily correlated with in-links.

Tables 2 and 3 in the Appendix show the top 20 blogs ranked by conversation mass vs. in-links. We found that the top 9 blogs by in-link count are in the top 20 by conversation mass. However, the reverse does not hold. Conversation mass surfaces important blogs like IMAO [3] and RadioEqualizer [4] that are buried by the in-link count metric. RadioEqualizer, which ranked at 53 in the in-link counts, reached rank 6 in conversation mass because of its leadership role in conversations about the Air America scandal in the summer of 2005.
This is interesting because the principal metric used to rank blogs has been inlink count (for example see the Technorati Top 100 [5] or BlogPulse's Daily Top Blogs [6]. One could argue that the the conversation mass metric is a better proxy for measuring influence. This metric captures the mass of the total conversation generated by a blogger, while number of inlinks captures only direct responses to the blogger's posts.

7 Conclusion

We have made several observations on what sort of features best characterize blogs in a network. Furthermore, we have provided an interesting look into how blogs evolve over time. We made some observations about cascade types. First, we note that the cascade types that blogs participate may suggest to which community it belongs ('humor', 'conservative', etc., see Observation 1) . Second, the number of trivial (singleton) cascades that a blog uses is a major indicator of cascade type (see Observation 2).

Next, we characterize blogs based on their general network characteristics, and observed that blogs tend to have posts that cluster together with respect to post features (Observation 3).
Next, we observed how blogs behave over time. The time behavior of blogs is bursty (Observations 4, 5). We propose a successful measure for the burstiness, the *bias factor*, which is related to the Hurst exponent of chaotic time series.

[3] http://www.imao.us
[4] http://radioequalizer.blogspot.com
[5] http://technorati.com/pop/blogs/
[6] http://www.blogpulse.com/links.html

We contribute several tools for analysis, which can serve as a basis for a general profile of a blog.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace., 2005.

[3] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, pages 509–512, 1999.

[4] A.-L. Barabasi. *Linked: The New Science of Networks*. Perseus Publishing, 1st edition, May 2002.

[5] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change in informational cascades. *Journal of Political Economy*, 100(5):992–1026, October 1992.

[6] M. Crovella and A. Bestavros. Self-similarity in world wide web traffic, evidence and possible causes. *Sigmetrics*, pages 160–169, 1996.

[7] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. In *SIGCOMM*, pages 251–262, 1999.

[8] A. T. Fiore, S. L. Tiernan, and M. A. Smith. Observed behavior and perceived value of authors in usenet newsgroups: bridging the gap. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 323–330, New York, NY, USA, 2002. ACM Press.

[9] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[10] M. Granovetter. Threshold models of collective behavior. *Am. Journal of Sociology*, 83(6):1420–1443, 1978.

[11] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04*, 2004.

[12] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03*, 2003.

[14] F. Korn, H. Jagadish, and C. Faloutsos. Efficiently supporting ad hoc queries in large datasets of time sequences. *ACM SIGMOD*, pages 289–300, May 13-15 1997.

[15] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03*, pages 568–576. ACM Press, 2003.

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

[17] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, New York, NY, USA, 2006. ACM Press.

[18] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model, October 2006.

[19] J. Leskovec, A. Singh, and J. Kleinberg. Patterns of influence in a recommendation network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[20] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing, 2002.

[21] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA, 1987.

[22] M. Schroeder. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. W.H. Freeman and Company, New York, 1991.

[23] C. E. Shannon and W. Weaver. *Mathematical Theory of Communication*. University of Illinois Press, 1963.

[24] A. Vazquez, J. G. Oliveira, Z. Dezso, K. I. Goh, I. Kondor, and A. L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73:036127, 2006.

[25] M. Wang, T. Madhyastha, N. H. Chang, S. Papadimitriou, and C. Faloutsos. Data mining meets performance evaluation: Fast algorithms for modeling bursty traffic. *ICDE*, Feb. 2002.

[26] D. J. Watts. A simple model of global cascades on random networks. In *PNAS*, 2002.

[27] G. Zipf. *Human Behavior and Principle of Least Effort: An Introduction to Human Ecology*. Addison Wesley, Cambridge, Massachusetts, 1949.

A Blogs ranked by conversation mass and in-link

Here we show the different rankings for blogs based on conversation mass and number of in-links.

Rank	Blog Url
1	michellemalkin.com
2	www.boingboing.net
*3	www.imao.us (75)
4	www.captainsquartersblog.com/mt
5	instapundit.com
6	radioequalizer.blogspot.com (53)
7	powerlineblog.com
8	www.waxy.org/links
9	www.washingtonmonthly.com
10	www.kottke.org/remainder
11	www.patriotdaily.com
*12	junkyardblog.net (34)
*13	mypetjawa.mu.nu (42)
*14	www.alternet.org/peek (58)
15	www.dailykos.com
16	wizbangblog.com
*17	digbysblog.blogspot.com (27)
18	stevegilliard.blogspot.com
* 19	drsanity.blogspot.com (84)
* 20	www.blackfive.net/main (67)

Table 2: Top 20 blogs according to conversation mass. The number in the parenthesis gives the rank of a blog using the number of the in-links

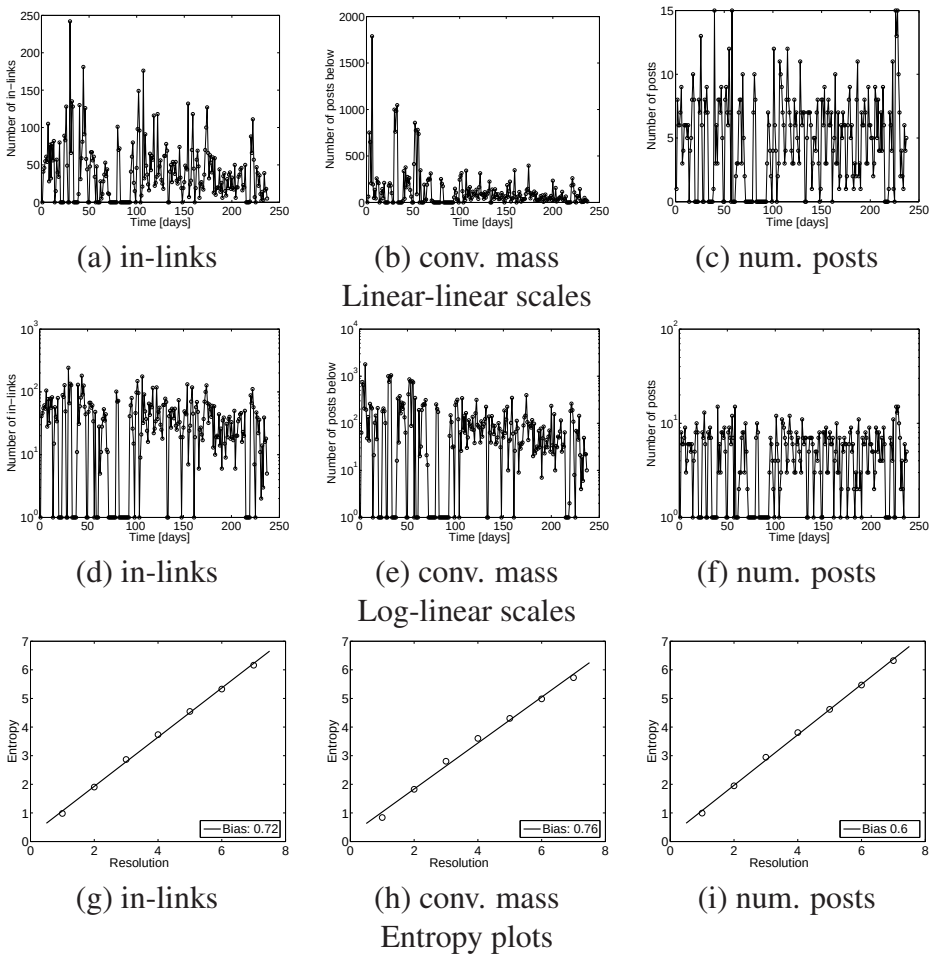
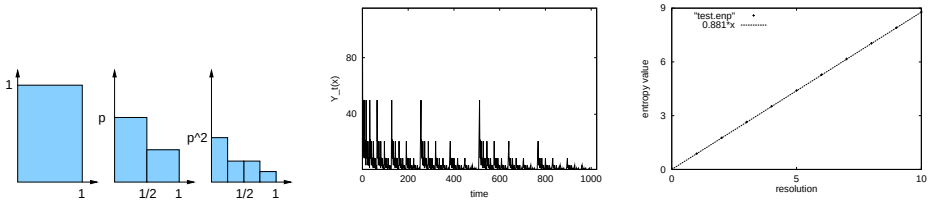
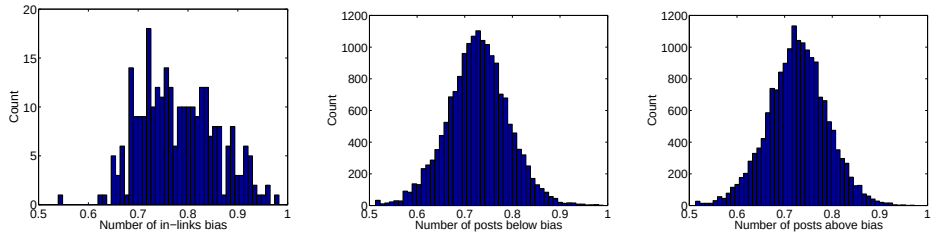


Figure 5: Time plots are bursty: in-links, conversation mass and number of posts, over time, for the `www.MichelleMalkin.com` blog. Top row: linear-linear axis; middle row: the 'y' axis is logarithmic. Bottom row shows the *entropy plots (*see text - entropy versus resolution r'*)*: they are all linear, which means that the time sequences are self-similar



(a) generation (b) synthetic data (c) entropy plot

Figure 6: Illustration of the *b-model*: *(*a*)* the recursive 80-20 procedure in its first three iterations *(*b*)* the generated synthetic activity *(eg., number of posts, over time*) *(*c*)* its *entropy plot (*entropy versus resolution - see text*)* Because the synthetic input traffic is self-similar, the entropy plot is linear, that is, *scale free*. Its slope is 0.881, much different than 1.0, which would be the uniform distribution *(*50-50*)*



(a) inLinks (b) downPosts (c) upPosts

Figure 7: Distribution of bias factors b of the blog time-sequences. Count, versus value of b, for *(*a*)* in-links over time, *(*b*)* number of posts above *(*upward conversation mass*)*, and *(*c*)* number of posts below *(*downward conversation mass*)*. *The other measures had similar behavior, and are omitted for brevity. Notice that the vast majority of are way above 0.5* (*uniform*)*, and closer to* b*=0.8 and 0.9*

Rank	Blog Url
1	www.boingboing.net
2	michellemalkin.com
3	instapundit.com
4	www.waxy.org/links
5	www.kottke.org/remainder
6	www.patriotdaily.com
7	www.captainsquartersblog.com/mt
8	powerlineblog.com
9	www.washingtonmonthly.com
*10	peteashton.com (30)
*11	www.gizmodo.com (35)
*12	www.eyebeam.org/reblog (33)
*13	billmon.org (31)
14	www.dailykos.com
*15	jeremy.zawodny.com/linkblog (50)
16	stevegilliard.blogspot.com
*17	www.sizemore.co.uk/blogmore.html (44)
*18	atrios.blogspot.com (26)
*19	www.juancole.com (51)
20	wizbangblog.com

Table 3: Top 20 blogs according to in-links. Number in the parentheses gives the rank by the blog conversation mass

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Carnegie Mellon