Modeling Blog Dynamics [*]

Michaela Götz [†]
Cornell University
goetz@cs.cornell.edu

Jure Leskovec [†]
Cornell/Stanford University
jure@cs.stanford.edu

Mary McGlohon
Carnegie Mellon University
mmcgloho@cs.cmu.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

Abstract

How do blogs produce posts? What local, underlying mechanisms lead to the bursty temporal behaviors observed in blog networks? Earlier work analyzed network patterns of blogs and found that blog behavior is bursty and often follows power laws in both topological and temporal characteristics. However, no intuitive and realistic model has yet been introduced, that can lead to such patterns.

This is exactly the focus of this work. We propose a generative model that uses simple and intuitive principles for each individual blog, and yet it is able to produce the temporal characteristics of the blogosphere together with global topological network patterns, like power-laws for degree distributions, for inter-posting times, and several more. Our model $\mathcal{ZC}$ uses a novel 'zero-crossing' approach based on a random walk, combined with other powerful ideas like exploration and exploitation. This makes it the first model to simultaneously model the topology *and* temporal dynamics of the blogosphere. We validate our model with experiments on a large collection of 45,000 blogs and 2.2 million posts.

1 Introduction

How do blogs (web-logs) initiate posts and link each other? Is there an intuitive model that produces these observed behaviors in the blogosphere? Blogs play a significant role in information dissemination, and here we seek to understand how patterns in blogosphere behavior arise from individual behaviors of blogs.

Blogs are web sites that are updated regularly, often in a journal style. Each update (or *post*) allows readers to make comments, as well as direct links to the readers' own blogs. The interaction between blogs can be viewed as a network of hyper-linked and timestamped posts, called "blogosphere".

Due to their timely and accessible nature, blogs have created a powerful social phenomenon, with blog discussions often influencing the mass media and public opinion (Adamic and Glance 2005), and the marketing industry. Blogosphere has experienced an explosive growth of two orders of magnitude in 3 years reaching about 50 million blogs in Aug 2006 (Sifry 2006).

Blogs exhibit community structure and temporal dynamic aspects, which makes them a richer domain of study than static web pages (Dezsö et al. 2006). Earlier work has found surprising patterns in blog dynamics: there are unexpected power laws in the popularity of blogs and the distribution of blog sizes, and self-similar (and bursty) patterns in the blog activity. Our goal is to understand the mechanisms in individual blogs that generate these patterns.

Applications of our work include modeling blog popularity and information diffusion in the blogosphere. Our model can be used to generate input to influence maximization algorithms (Kempe, Kleinberg, and Tardos 2003; Leskovec et al. 2007a) which can be applied in viral marketing campaigns and web advertising.

Our contribution is the proposed zero-crossing ($\mathcal{ZC}$) model: it is simple and intuitive while it requires no tuning of parameters. Yet it successfully matches observed power law distributions, temporal burstiness. The $\mathcal{ZC}$ model is first to jointly model the temporal dynamics and the structural properties of the blogosphere.

Formulating an appropriate model is vital for understanding how blogs interact for extrapolation and forecasting purposes. Moreover, our findings in blog dynamics could help us form hypotheses about the general flow of information, which may have applications in marketing or epidemiology.

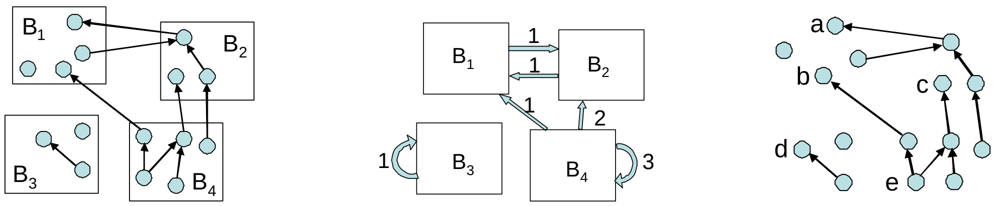
The rest of this paper is organized as follows. Section 2 gives the literature survey. Section 3 states the patterns occurring in the blogosphere. Section 4 describes our proposed model. We experimentally validate our model in Section 5 and conclude in Section 6.

2 Related Work

As mining and modeling of blogs and related social networks has attracted a lot of interest, we focus on surveying models for the blogosphere and networks in general, see (Jensen and Neville 2002) for an extensive survey on learning models in networks.

[*] This material is based upon work supported by the National Science Foundation under Grants No. IIS0808661 This work is also partially supported by an IBM Faculty Award, a Yahoo Research Alliance Gift, a SPRINT gift, with additional funding from Intel, NTT and Hewlett-Packard. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties.

[†] work performed at CMU



(a) Blogosphere (b) Blog network (c) Post network

Figure 1: A graphical representation 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. We create two networks: a blog network (b) of citations (links) between blogs, and a post network (c) with time stamped links between blog posts.

Blog models. There has been extensive work modeling different characteristics of blogs. In (Venolia) a large blogging community was studied and a model for blog mortality was presented. The authors of (Kumar et al. 2003) argue that a random linking behavior cannot explain the dynamics of the community structure. Instead of linking randomly the authors of (Karandikar et al. 2008) applied the preferential attachment rule (see below) to create realistic links (w.r.t. degree and component distribution).

A different line of work models the information propagation. The authors of (Adar and Adamic 2005) discovered patterns in linking behavior and used a support vector machine. Epidemiological models have also been used in this context. See (Bailey 1975) for details on such models, like the "SIS" (susceptible–infected–susceptible) and "SIR" (susceptible–infected–removed) ones. In (Gruhl et al. 2004) an SIR-based model of information propagation with respect to topics was presented. Later the authors of (Leskovec et al. 2007b) presented an SIS-based model producing realistic cascades, i.e., graphs of information propagation.

Related models. There have also been several models for human behavior in other dynamic environments that may serve as inspiration for a model for blog behavior. One line of research models the structure of networks. Prevalent is the "preferential attachment" rule (Barabási and Albert 1999) and variations (Chung and Lu 2006; Pennock et al. 2002). In (Pathak, Mane, and Srivastava 2006) a socio-cognitive network based on email communication was modeled.

Another line of research models temporal aspects, for example the time between answering two consecutive emails at a single user which follows a power law with exponent -1.5, see (Barabasi 2005; Vazquez et al. 2006). Similarly, in (Kleinberg 2002) a weighted 2-state Markov Chain based model of inter-arrival times of emails was introduced. However, while the models are intuitive, they fail to generate temporal bursty behavior is found in blogs.

Basically, none of the above models is able to match as many properties of real blogs as our upcoming $\mathcal{ZC}$ model which models both temporal and topological characteristics.

3 Background and Problem Definition

Next we describe patterns that we would like our model to produce. We distinguish two types of patterns: *topological* and *temporal*. Topological patterns refer to structural patterns of the blog network, like degree distribution, while temporal patterns involve time, like uniformity/burstiness measures of the number of posts per unit time.

First we describe known patterns, and then show a pattern we discovered in the course of this work. To our knowledge this work presents the most complete model that matches the largest number of patterns that we have seen in the literature so far; earlier models typically focus on modeling the emergence of only one of these patterns. Modeling more than a single characteristic is important as models become more realistic, more powerful and more widely applicable.

Old Patterns

In our earlier work (Leskovec et al. 2007b; McGlohon et al. 2007) that forms the background of this paper we analyzed a data set of 45,000 blogs and approximately 2.2 million posts. We defined two networks of interest: the *Blog network* (Fig. 1(b)) and the *Post network* (Fig. 1(c)). In the Post network, nodes represent individual posts and each edge represents a hyper-link from one post to another, earlier post. Edges are labeled with time-stamps of the link occurrence (that is, the time at which the source of the link, the referring post, was written). Posts across blogs that participate in the same discussion can be viewed as being part of the same conversation tree (i.e., *cascade*). In the Blog network, nodes represent blogs. A directed edge from blog B_1 to B_2 means that at some point in time, a post at B_1 linked to a post at B_2 (Fig. 1). Studying these two networks, we pointed out several interesting patterns:

BID (topological) The probability density function (PDF) of the Blog In-Degree follows a power law.

PID (topological) The PDF of the Post In-Degree follows a power law.

SCT (topological) The PDF of the Size of non-trivial Conversation Trees in the post network follows a power law.

PP (temporal) The Popularity of Posts, i.e., the number of in-links of a post, versus post age, drops with a power law with exponent -1.6.

IFD (temporal) The activity of blogs is bursty and self-similar. The *Information Fractal Dimensions* are in large part between 0.72 and 0.88. We explain the concept of information fractal dimension in the next section.

For example, Figure 4 shows the topological power laws and their exponents (-1.7 for BID, -2.15 for PID, -1.97 for SCT). Temporal patterns are shown in Figure 5.

New Pattern: Inter-Posting-Time

Through further analysis we discovered the following temporal pattern (see Figure 5(b)).

IPT (temporal) The PDF of the Inter-Posting-Time follows a power law of exponent -2.7. The inter-posting time is defined as the time between two consecutive posts of the same blogger.

Definition: Information Fractal Dimension

What does it mean that the posting activity of a blog is bursty and self-similar? A common measure of burstiness is the *fractal dimension* (or *intrinsic dimension*). Here we use a variant called the *information fractal dimension* (Wang et al. 2002). Intuitively, the fractal dimension of a cloud of points (i.e., time-stamps on the time-line) is roughly the degrees of freedom: A cloud of 3-d points, all lying on a 2-d plane, has intrinsic dimensionality $f = 2$. It is surprising that real and synthetic clouds of points often have fractional intrinsic dimensionality: E.g., Cantor dust ("delete the middle-third") (Schroeder 1991) has fractal dimension $f = 0.63$.

The "information fractal dimension" is defined as the slope of the *entropy-plot* (Wang et al. 2002). The plot shows how the entropy changes as a function of the resolution (e.g., Fig 5(a)). In more detail, consider a set $\mathcal{T}$ of n time-stamps t_1, ..., t_n, in a time interval of duration T of time-ticks. In our case these could be the time-stamps of the posts we are interested in, and can be envisioned as 1-dimensional points.

The entropy $H(W)$ at window size W is defined as follows. Let $n_{i,W}$ be the number of events (e.g., posts) at interval i, after we have divided our duration T into disjoint, consecutive windows each of duration W. Let $p_{i,W}$ be the fraction of events that fall into the i-th such interval — clearly $p_{i,W} = n_{i,W}/n$. Then we define $H(W)$ as

$$H(W) = -\sum_i p_{i,W} \log_2(p_{i,W})$$

The entropy plot is defined as the plot of $H(W)$ versus $\log_2(W)$.

If a process is self-similar, its entropy plot is linear. The intrinsic ("fractal") dimension f is then defined as the slope of the entropy plot $f = \frac{\partial H(W)}{\partial(\log(W))}$. The value of f then indicates how bursty the activity is – the lower, the burstier.

Figure 5(a) shows the entropy plots for two example blogs; a real one (about politics: MichelleMalkin.com), and a synthetic blog generated by our zero-crossing ($\mathcal{ZC}$) model. The time-interval covers $T = 2^7$ time-ticks (for the real blog) and $T = 2^{15}$ time-ticks (for the synthetic blog). Every time-stamp corresponds to a post that is published by the blog at that time-tick. With the entropy plots we measure how the posts are distributed on the time-interval.

If the posting activity was uniform, the information fractal dimension would be $f = 1$ (one degree of freedom); if the activity was concentrated (i.e, all posts happen on exactly the same time), the fractal dimension would be zero (the dimensionality of a point). In the real data, the dimensionality of the distribution of time-stamps is somewhere in-between.

Problem Definition

We want a natural model that matches the above mentioned statistical patterns. Semi-formally, our goal is the following: *to devise a set of simple principles or local rules that each blogger would follow, so that these principles lead to emerging, macroscopic behavior that matches the patterns we listed above (BID, PID, etc.)* Notice that this is an ambitious goal. Previous models for blog behavior mostly focused on a single topological aspect of the blogosphere. On the other hand, our model here is different as it models both the temporal aspects as well as topological aspects.

Alternative Models

It is a challenging to come up with a set of principles that produces when followed by each individual blogger, which give rise to the global temporal and topological patterns and power laws that we observe in the real data. Most textbook type behaviors, like Markov-chain based ones, do not lead to power laws, but exponential behavior.

Moreover, patterns are difficult to create naturally. There are only a few models known that create self-similar and bursty behavior. One of the models is the zero-crossing model which we embedded in our model $\mathcal{ZC}$. Another model is the "b"-model (Wang et al. 2002), where a time interval is divided in two intervals and a constant fraction "b" of the activity is assigned to one interval and the remaining fraction 1-"b" of the activity is assigned to the other interval. When proceeding recursively, the activity is spread burstily and self-similarly over the whole interval. However, this model does not comply with our intuition of the blogger behavior. We cannot imagine that a blogger plans his blogging activity for a whole time-interval in advance or that he takes his whole blogging past into account when deciding whether to blog at a time-tick or not.

The blog models discussed in related work mainly focused on information propagation, whereas our focus are topological and temporal patterns. All previous email communication models mainly focused on the time between two incoming emails and the time between answering emails, which motivated us to analyze the inter-posting time IPT. However, we will not focus on the *single aspect* (inter-arrival time or inter-sending time) like the models proposed for the email traffic do. Instead we simultaneously model topological and temporal behavior. Furthermore, some models are *not natural*, such as the growth function in (Huberman and Adamic 1999; Karandikar et al. 2008), or the exponential distribution in (Kleinberg 2002).

Moreover, some models *need assumptions*, such as a constant rate of answering emails (Vazquez et al. 2006).

Comparison to other models. Our model puts together two very different aspects of the blogosphere, time and topology, properties that are much more difficult to model jointly than when considered separately. As existing models usually consider modeling single aspect of the blogosphere such as the mortality of blogs or the information propagation there is no natural model to compare our model to. However, in order to have a baseline comparison, we devised a nontrivial model based on conventional wisdom of exponential post

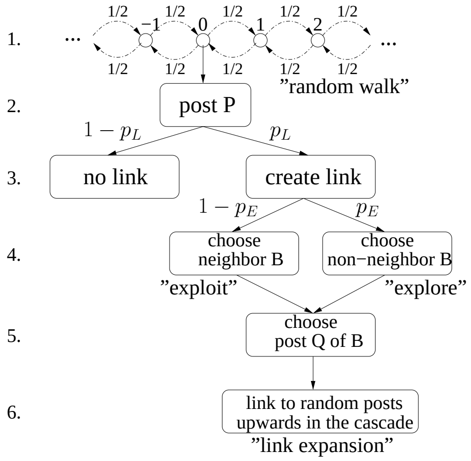


Figure 2: Our zero-crossing model $\mathcal{ZC}$. Each blog behaves according to this model. Numbers correspond to the steps of our $\mathcal{ZC}$ generative model.

inter-arrival times (Kleinberg 2002) and "rich get richer" linking behavior. We refer to it as the $\mathcal{EXP}$ model which we define as follows. The inter-posting times for each blog are sampled from an exponential distribution with parameter λ. A blog then creates a post and links to another post that is chosen by the "preferential attachment" rule (Barabási and Albert 1999): the probability of linking to a post is proportional to its current in-degree, which is a measure of its current popularity.

4 Proposed Model: Zero-Crossing $\mathcal{ZC}$

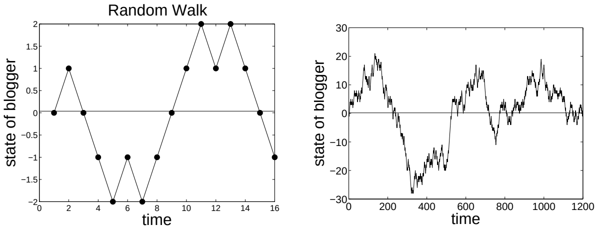


Figure 3: Random walk over the states of a blogger. Left, a blogger posts at times 1, 3, 9, 15 when the random walk crosses horizontal axis which gives inter-posting times 2, 6, 6. Right, a longer walk demonstrates the burstiness.

Next we describe our zero-crossing model ($\mathcal{ZC}$) based on a random walk on a line, which is sketched in Figure 2.

Our model involves three major mechanisms, each handling one aspect of the dynamics of the blogosphere:

- *WHEN*: When would a blogger write a post? We propose a model based on zero-crossing of a random walk on a discrete line.
- *WHAT*: Once a blogger has decided to blog, which other blogs (if any) will he choose to read, and which posts inside those chosen blogs will he choose to cite? Our idea here is related to the "exploit and explore" strategy: usually, the blogger will choose one of the blogs he has chosen in the past ("exploit"), but occasionally he will read a completely new blog ("explore").
- *FOLLOW-UP*: Once a blogger decides to cite post Q, he may follow up on it, and also cite some of the posts that Q is citing; the blogger may do that recursively. We will refer to this mechanism as *link expansion*.

Next we describe the details of each of the mechanisms.

"WHEN" and "random walk": The heart of our model is a natural way that will generate power-laws and self-similarity in temporal posting activity.

We propose the following mechanism: The blogger does a random walk on a line, and decides to post whenever he is at state **0** (e.g., at his computer). At each time tick, a blogger is in a state represented by an integer. There are two possible transitions: with equal probability the blogger either adds or subtracts 1 from his current state. Blogger publishes a post when his state is **0**. In that sense, the state of a blogger describes how far away he is from his computer (or equivalently, how far he is mentally away from the blogging mode). The idea is that random events may distract him to some other, nearby state; if there are too many successive distractions away from state **0**, the blogger will be away from his computer for a long time. This mechanism *provably* generates bursty blogging activity: the blogging time-stamps are exactly the zero-crossings of a random walk (Brownian motion), and it is known that their intrinsic ("fractal") dimension is $f = 0.5$ (Mandelbrot 1982; Schroeder 1991). See Fig. 3 for examples of random walks.

Random walks have also been considered to model and explain how human make decisions in uncertain environments, for instance see (Busemeyer and Townsend 1993).

"WHAT" and "exploit and explore": Once the blogger is ready to post, he may choose to initialize a new conversation tree (with probability $1 - p_L$), i.e., a new post without any outlinks that other can then cite to create new information cascade. The interesting modeling aspects arise in the opposite case, when the blogger decides to comment on some other posts and join to an existing conversation tree (information cascade).

How does he choose a posts to comment on? We propose the following mechanism, which reflects how humans act: the chosen post will belong to one of his favorite blogs. However, once in a while, the blogger may want to cite a post on a completely new blog. Thus, first the blogger decides whether to pick a post of a neighbor ('favorite blog') or a post of a non-neighbor in the blog network. Among the neighboring blogs, possible candidates are blogs that have published a post since his last visit. He prefers candidates that he preferred in the past, e.g. he chooses a blog proportionately to the number of past links he has made to that blog. We call this the "exploit" mode, where blogger visits favorite blogs that he found valuable/interesting in the past.

In the opposite case, with probability p_E, the blogger goes into the "explore" mode and chooses a blog he has never linked to before. In that case, he trusts the taste of the majority and chooses a blog B proportionally with the total number in-links of B times the number of posts of B. We expect a rich-get-richer setting, because blogs with many in-links probably have higher quality and/or better word-of-mouth ratings, and thus will naturally attract attention of bloggers.

After choosing a blog, the blogger has to determine on which post to comment. He therefore judges the posts based on their recency and their popularity, i.e., the probability of linking to a post is proportional to the ratio of the number of in-links and the time since the publication of the post.

"FOLLOW-UP" and "link expansion": Now, our blogger can publish his post with a link to the chosen post. He will consider to link to other posts that participated in the same conversation tree, in the same way that scientific papers point to an earlier article A, and often point to the citations of A, and so on recursively. Posts that are many hops away from the chosen post are less likely to be linked: for each post and each path p from the chosen post to that post he flips a biased coin and with probability $p_{LE}^{|p|}$ he links it.

Notice that our proposed $\mathcal{ZC}$ model heavily relies on how the information flows through the blogosphere. We exploit this both in a topological sense to model how bloggers create links and in a temporal sense to model the dynamics at which new posts are being written.

This completes the description of our artificial blogger. After that, our the blogger transitions the state, and it continues with simulating the next blogger, in a round-robin fashion. Notice that all the three major steps in our blogger model have very simple, local behavior, with no sophisticated distributions or constraints to guide our blogger. Yet, as we show next, this simple model, repeated over all bloggers, leads to emerging behavior that matches the properties and patterns found on the real blogosphere.

Formal Description of our $\mathcal{ZC}$ Model

Each blogger has 3 parameters: p_L (prob. of a post creating an out-link), p_E (prob. of exploration mode), and p_{LE} (prob. of expanding a link). All blogs start at position **0** and publish a post in the first round. In each next round each blog A follows the 6 steps of Fig. 2 which we describe next:

1. Change state: With probability 1/2 add one to current state of A, and with probability 1/2 subtract one A's state.

2. Create post: If A's current state is not **0** then stop else continue with next step.

3. Initiate cascade: A creates a post P. With probability $1 - p_L$, A initializes a new conversation tree (P has no out-links) and stop else continue with next step.

4. Choose mode: With probability p_E blog A is in "exploration" mode and with $1 - p_E$ it is in "exploitation" mode.

4.1. "exploitation" mode: Let $N(A)$ be the set of neighboring blogs, blogs A previously linked to. Then the probability of A choosing a neighboring blog B is: $\Pr[A \text{ chooses } B] \propto \#\text{links}(A \to B)$

4.2: "exploration" mode: A chooses a non-neighbor blog. Let $\bar{N}(A)$ be the set of blogs with no in-links from A. The probability of choosing a non-neighbor B is: $\Pr[A \text{ chooses } B] \propto (\#\text{inlinks}(B) + 1)(\#\text{posts}(B) + 1)$

5. Choose post: The probability of choosing a post Q in blog B is: $\Pr[A \text{ chooses } Q] \propto \frac{\#\text{inlinks}(Q)+1}{\#\text{rounds passed since publication}+1}$. A creates a link from its post P to the post Q of blog B.

6. Link Expansion: For each post R reachable from post P, for each path p from P to R with probability $p_{LE}^{|p|}$ create a link from post P to post R.

Analysis of our Model $\mathcal{ZC}$

Theorem 1 *The inter-posting times in our model $\mathcal{ZC}$ follow a power law distribution with exponent -1.5.*

***Proof* 1 (Sketch)** *(Newman 2005)* We first note that the probability of posting at time t in our model (denoted by $u_{t'}$) is zero for odd t' and $2^{-t'}\binom{t'}{t'/2}$ otherwise. We can relate $u_{t'}$ (for even $t' > 0$) to the probability of the inter-posting being t (denoted by p_t) as follows:

$$u_{t'} = \sum_{1 \le t \le t'/2} p_{2t} u_{t'-2t}$$

Solving for p_{2t} we obtain $p_{2t} = \frac{\binom{2t}{t}}{(2t-1)2^{2t}}$. Using Sterlings formula in a limit analysis ($t \to \infty$) we obtain the result:

$$p_t \propto t^{-3/2}$$

Theorem 2 *The blogging activity in our $\mathcal{ZC}$ Model is self-similar and bursty.*

***Proof* 2** The intrinsic ("fractal") dimension of the zero-crossings of Brownian motion is is f = 0.5, see for example (Mandelbrot 1982; Schroeder 1991). This result extends to our random walk which is a discrete version of Brownian motion.

5 Experiments – Model Validation

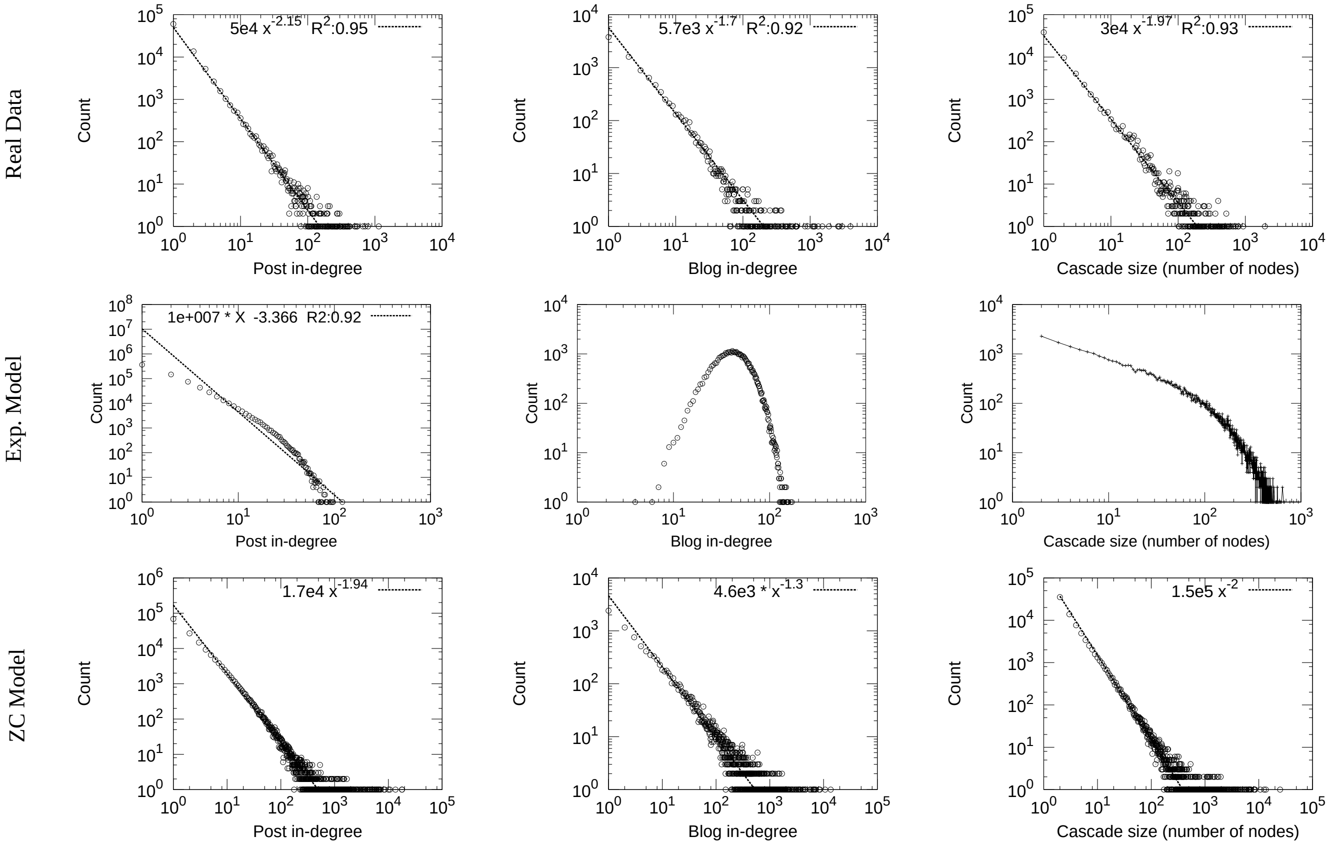
Experimental Setup

We validate our $\mathcal{ZC}$ model on a set of 45,000 blogs with 2.2 million posts from August and September 2005 (Leskovec et al. 2007b). We started with a set of 50 million blogs (Glance et al. 2005) but since most of them do not actively participate in the blogosphere, we biased our dataset set towards active blogs.[1] We represent the data as Blog network and as Post network (see Figure 1), where edges are labeled with a timestamp.

Validation

We validate our model $\mathcal{ZC}$ through the topological and temporal properties and patterns found in the real blogosphere. We compare distributions of properties in the real data with those in the synthetic data produced by our models. For comparison we also employ the baseline $\mathcal{EXP}$ model. We consider a model to be good if it intrinsically produces patterns and properties similar to those found in the real data. Note that statistical properties of conversations and blog behavior intrinsically emerge from the model and were not in any way "forced".

[1] There are two possible extensions to our model that account for inactive blogs. First we could sample the lifespan of a blog as done in (Venolia). Second we could initiate the state of some blogs to be far away from **0**.



(a) PID: Post In-Degree (b) BID: Blog In-Degree (c) SCT: Size of Cascades

Figure 4: Topological patterns of the blogosphere. Top: real blogosphere; Middle: $\mathcal{EXP}$ model; Bottom: blogosphere as modeled by the $\mathcal{ZC}$ model. Notice $\mathcal{ZC}$ model outperforms $\mathcal{EXP}$ model and matches the properties of real blogosphere.

Setting the Parameter Values

In the $\mathcal{ZC}$ model for each blog we chose the parameters p_L, p_E and p_{LE} independently and uniformly at random in $[0, 1]$. So, in $\mathcal{ZC}$ model there are no parameters to set or tune. In order to achieve a good basis of comparison between the real data and the synthetic data, we chose the number of blogs in the simulation to be 45,000 and and run it till 2.2 million posts are created. For $\mathcal{ZC}$ model there are no parameters to set, while for the $\mathcal{EXP}$ model we choose the parameter λ, such that on average a time unit corresponds to an hour in the real data.

Topological Patterns - Blogosphere

Figure 4 shows that the power laws in the distribution of the BID the PID and the SCT found by (Leskovec et al. 2007b) are matched closely by our $\mathcal{ZC}$ model. Not only $\mathcal{ZC}$ matches the shape perfectly, but it also matches the power law exponents well: -1.94 versus -2.15 for the BID in Fig. 4(a); -1.3 versus -1.7 for the PID in Fig. 4(b); and -2 versus -1.97 for the SCT in Fig. 4(c). In contrast $\mathcal{EXP}$ model only somewhat mimics the PID power law.[2]

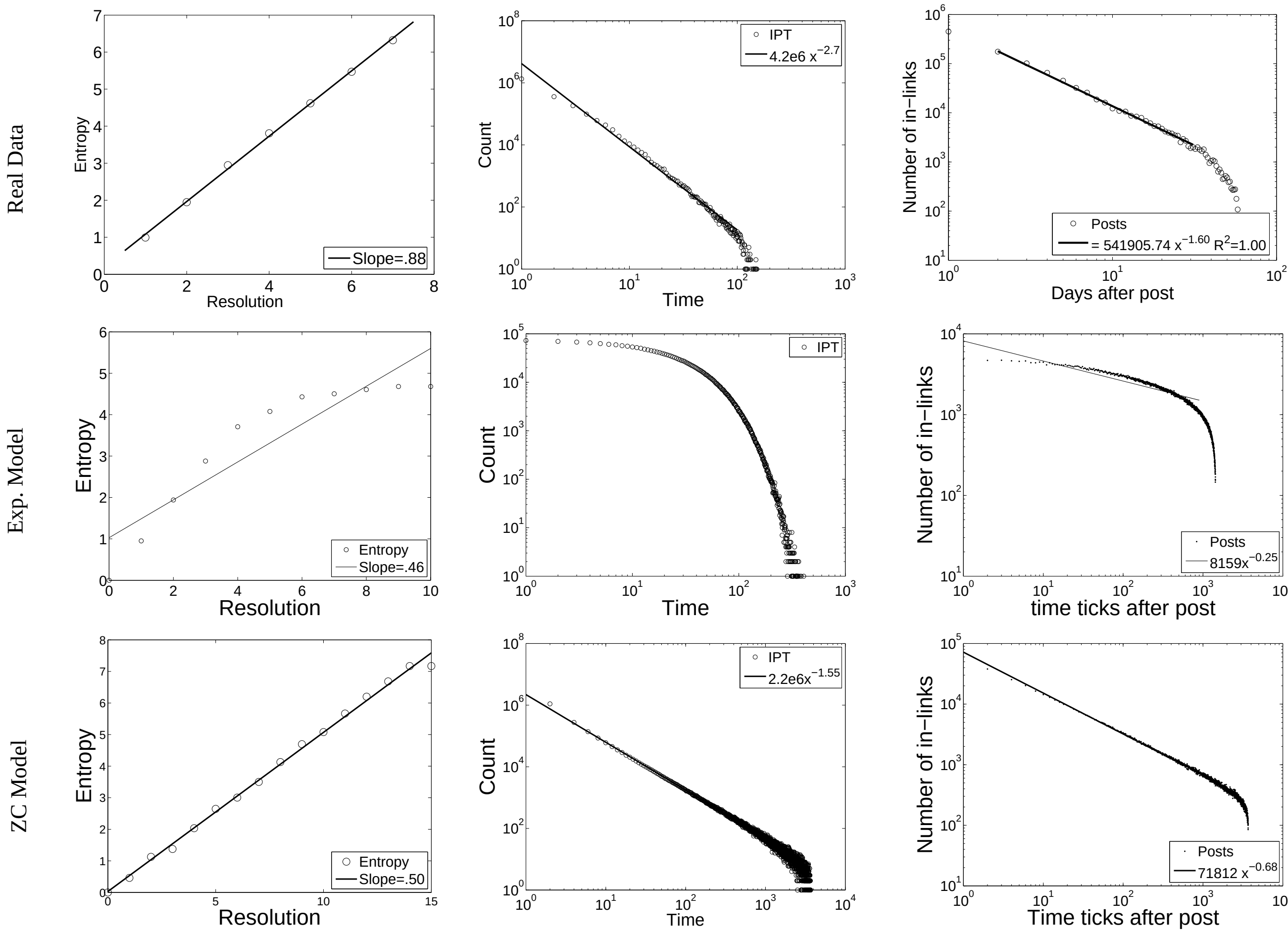
Where do the power laws come from in our model $\mathcal{ZC}$? The power laws of the in-degree distributions can be explained by the fact that a blog A in the "exploration" mode chooses another blog B in order to link to a post published by B based on the number of B's in-links, which causes a rich-get-richer phenomenon. This phenomenon leads to a power law distribution. Similarly, a blog A publishing a post chooses another post P to create a link to P based on number of P's in-links. Again, the resulting rich-get-richer phenomenon leads to a power law distribution.

Moreover, the $\mathcal{ZC}$ model also matches the power law of the distribution of the cascade sizes (SCT) which is more surprising. Our model $\mathcal{ZC}$ is the first blog model that matches this power law. The power law exponents are almost the same (-2 versus -1.97).

Temporal Patterns

Information Fractal Dimension (IFD): From entropy plots of (McGlohon et al. 2007) we observe that the activity of most blogs is self-similar and bursty. Our model $\mathcal{ZC}$ also creates bursty and self-similar activity, as can be seen in Fig. 5(a). The entropy plots plot the entropy versus resolution, that is $H(W)$, vs. $\log_2(W)$. The plots of the real data

[2] The power law comes out more clearly if the model is run for longer time.



(a) IFD: Entropy Plot (b) IPT: Inter-Posting Times (c) PP: Popularity over Time

Figure 5: Temporal patterns of the blogosphere. Top: real blogosphere; Middle: $\mathcal{EXP}$ model; Bottom: blogosphere as modeled by the $\mathcal{ZC}$ model. Notice $\mathcal{ZC}$ model outperforms $\mathcal{EXP}$ model and matches the temporal characteristics of real blogosphere.

and the synthetic data generated by $\mathcal{ZC}$ model are both linear which implies that the activity is self-similar as discussed in section on information fractal dimension. Furthermore, both plots have a slope different from 1, which implies that the activity is bursty and not uniform. Similar plots can be found for most of the blogs (McGlohon et al. 2007) in the real data. In fact, our model provably creates self-similar and bursty activity, see Thm. 2. In contrast, the $\mathcal{EXP}$ model does not create self-similar activity (left middle plot of Fig. 5). Moreover, we can extend the $\mathcal{ZC}$ model to match the slope of the real data more accurate by modifying in an ad-hoc fashion the random walk into a more general form of Brownian motion, a.k.a., anomalous diffusion (Ding and Yang 1995).

Inter-Posting Times (IPT) A different though related approach to analyzing the temporal activity of a blog, is to focus on the inter-posting times (IPT), which is shown in figure Fig. 5(b) (in log-log scales). Our model $\mathcal{ZC}$ matches the shape of the power law distribution perfectly. In fact, the first return times (in our $\mathcal{ZC}$ model: the inter-posting times) follow a power law distribution with exponent -1.5, as we showed in Thm. 1. In contrast, the inter-posting times of the $\mathcal{EXP}$ model follow an exponential distribution.

Popularity of Posts over Time (PP) Another dynamic aspect of the blogosphere is the number of in-links a post published at time t obtains at time $t + \delta$. The plot basically measures how quickly does the popularity (number of on-links) of a post decay with its age. Fig. 5(c) depicts δ on the horizontal axis and it depicts the overall number of links that were created δ time-ticks after the publication of the post it links to on the vertical axis. Again, note that the power law discovered in (Leskovec et al. 2007b) is matched more closely by our model $\mathcal{ZC}$ than by the $\mathcal{EXP}$ model.

Where does this power law come from in our model $\mathcal{ZC}$? A blogger chooses a post of a blog by its recency and its number of in-links, that is, the probability is given by normalized ratio of number of in-links and the time difference since the publication of the post. Since a blog publishes at most one post per time-tick it follows that the PDF of the time differences that occur in that selection of posts is the time difference multiplied by the number of in-links of the corresponding post. Globally, a power law distribution of time differences emerges that matches the real data.

6 Conclusions

We presented a novel "zero-crossing" ($\mathcal{ZC}$) model for blog dynamics that naturally generates several of the patterns and power-laws that were observed in the structure and dynamics of the blogosphere. The model uses novel ideas, such as the zero crossings of a random walk and the "link expansion", and has the following desirable properties:

(a) It is *simple and intuitive*, mimicking simple rules that a human blogger would follow, which may lend some insight into other online human behaviors.

(b) It creates realistic blogospheres, matching all the *topological* patterns, namely the post in-degree (PID), the blog in-degree (BID), the cascade sizes (SCT) (see Fig. 4).

(c) $\mathcal{ZC}$ model matches *temporal* patterns, *burstiness* (IFD) and power laws in the inter-posting time (IPT), and the popularity of posts over time (PP) (see Fig. 5).

(d) Our model requires no magic parameters to set as we show that even random parameter settings give good results.

(e) We validate our model with experiments on a large collection of blogs (2.2 million posts), and we discover a new power law, governing the inter-post time distribution (IPT).

Our model can naturally be used to generate synthetic blogospheres for what-if scenarios, to explore and model blog dynamics for the purposes of information propagation, marketing, and advertising.

References

Adamic, L. A., and Glance, N. 2005. The political blogosphere and the 2004 U.S. election: divided they blog. In *LinkKDD*.

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Web Intelligence*, 207–214.

Bailey, N. 1975. *The Mathematical Theory of Infectious Diseases and its Applications*. London: Griffin.

Barabási, A.-L., and Albert, R. 1999. Emergence of scaling in random networks. *Science* 286:509–512.

Barabasi, A.-L. 2005. The origin of bursts and heavy tails in human dynamics. *Nature* 435:207.

Busemeyer, J. R., and Townsend, J. T. 1993. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* 100(3):432–459.

Chung, F., and Lu, L. 2006. *Complex Graphs and Networks*. American Mathematical Society.

Dezsö, Z.; Almaas, E.; Lukacs, A.; Racz, B.; Szakadat, I.; and Barabási, A.-L. 2006. Dynamics of information access on the web. *Phys Rev E* 73(6).

Ding, M., and Yang, W. 1995. Distribution of the first return time in fractional brownian motion and its application to the study of on-off intermittency. *Phys. Rev. E* 52(1):207–213.

Glance, N. S.; Hurst, M.; Nigam, K.; Siegler, M.; Stockton, R.; and Tomokiyo, T. 2005. Deriving marketing intelligence from online discussion. In *KDD*.

Gruhl, D.; Guha, R. V.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW*, 491–501.

Huberman, B. A., and Adamic, L. A. 1999. Growth dynamics of the world-wide web. *Nature* 399.

Jensen, D., and Neville, J. 2002. Data mining in social networks. In *National Academy of Sciences Symposium on Dynamic Social Network Analysis*.

Karandikar, A.; Java, A.; Joshi, A.; Finin, T.; Yesha, Y.; and Yesha, Y. 2008. Second Space: A Generative Model For The Blogosphere. In *ICWSM*. AAAI. Poster.

Kempe, D.; Kleinberg, J.; and Tardos, E. 2003. Maximizing the spread of influence through a social network. In *KDD*.

Kleinberg, J. 2002. Bursty and hierarchical structure in streams. In *KDD*.

Kumar, R.; Novak, J.; Raghavan, P.; and Tomkins, A. 2003. On the bursty evolution of blogspace. In *WWW*, 568–576.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007a. Cost-effective outbreak detection in networks. In *KDD*.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007b. Cascading behavior in large blog graphs.

Mandelbrot, B. B. 1982. *The Fractal Geometry of Nature*. W. H. Freeman.

McGlohon, M.; Leskovec, J.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Finding patterns in blog shapes and blog evolution. *ICWSM*.

Newman, M. E. J. 2005. Power laws, pareto distributions and zipf's law. *Contemporary Physics* 46:323.

Pathak, N.; Mane, S.; and Srivastava, J. 2006. Who thinks who knows who? Socio-cognitive analysis of email networks. In *ICDM*, 466–477.

Pennock, D. M.; Flake, G. W.; Lawrence, S.; Glover, E. J.; and Giles, C. L. 2002. Winners dont take all: Characterizing the competition for links on the web. In *Proceedings of the National Academy of Sciences*.

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

Sifry, D. 2006. State of the blogosphere. Technical report, Technorati.

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

Venolia, G. A matter of life or death: Modeling blog mortality. Technical report, Microsoft Research.

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