The Information Ecology of Social Media and Online Communities

Tim Finin, Anupam Joshi, Pranam Kolari, Akshay Java, Anubhav Kale, and Amit Karandikar

■ *Social media systems such as weblogs, photo- and link-sharing sites, wikis, and online forums are currently thought to produce up to one third of new web content. One thing that sets these "web 2.0" sites apart from traditional web pages and resources is that they are intertwined with other forms of networked data. Their standard hyperlinks are enriched by social networks, comments, trackbacks, advertisements, tags, RDF data, and metadata. We describe recent work on building systems that use models of the blogosphere to recognize spam blogs, find opinions on topics, identify communities of interest, derive trust relationships, and detect influential bloggers.*

Web-based social media systems such as blogs, wikis, media-sharing sites, and message forums have become an important new way to transmit information, engage in discussions, and form communities on the Internet. Their reach and impact is significant, with tens of millions of people providing content on a regular basis around the world. Recent estimates suggest that social media systems are responsible for as much as one third of new web content. Corporations, traditional media companies, governments, and nongovernmental organizations (NGOs) are working to understand how to adapt to them and use them effectively. Citizens, both young and old, are also discovering how social media technology can improve their lives and give them more voice in the world. We must better understand the information ecology of these new publication methods in order to make them and the information they provide more useful, trustworthy, and reliable.

The blogosphere is part of the web and therefore shares most of its general characteristics. It differs, however, in ways that affect how it should be modeled, analyzed, and exploited. The common model for the general web is as a directed graph of web pages with undifferentiated links between pages. The blogosphere has a much richer network structure in that there are more *types* of nodes that have more *kinds* of relations between them (figure 1). For example, the people who contribute to blogs and author blog posts form a social network with their peers, which can be induced by the links between blogs. The blogs themselves form a graph, with direct links to other blogs through *blog rolls* and indirect links through their posts. Blog posts are linked to their host blogs and typically to other blog posts and web resources as part of their content. A typical blog post has a set of comments that link back to people and blogs associated with them. Finally, the blogosphere trackback protocol generates implicit links between blog posts. Still more detail can be added by taking into account post tags and categories, syndication feeds, and semistructured metadata in the form of extensible markup language (XML) and resource description framework (RDF) content.

In this article, we discuss our ongoing research in modeling the blogosphere and extracting useful information from it. We begin by describing an overarching task of discovering which blogs and bloggers are most influential within a community or about a topic. Pursuing this task uncovers a number of problems that must be addressed, three of which we describe in more detail. The first is recognizing spam in the form of spam blogs (splogs) and spam comments. The second is developing more effective techniques to recognize the social structure of blog communities. The final one involves devising a better abstract model for the underlying blog network structure and how it evolves.

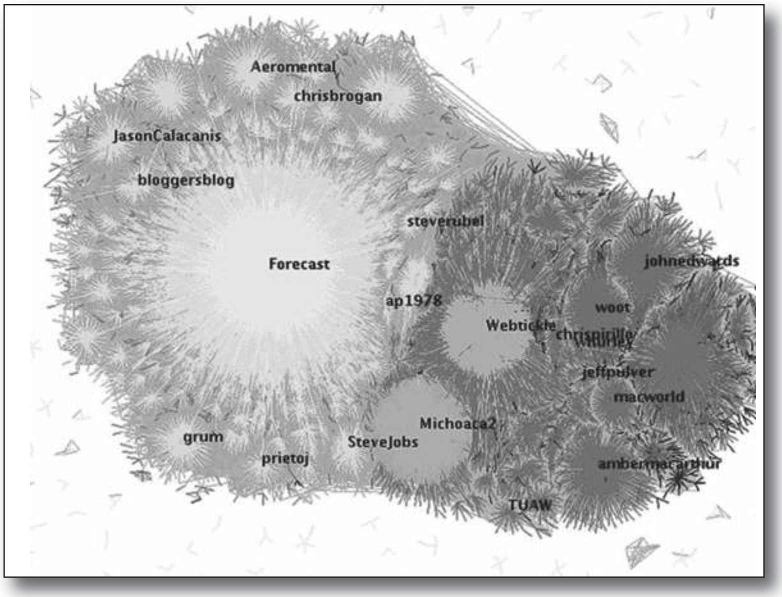


Figure 1. A Social Network Formed by Users of the Twitter Microblogging System.

Modeling influence and information flow in social media systems requires attention to many factors, including link structure, semantic metadata, named entity recognition, conversational structures, sentiment analysis, and topic classification.

1	www.talkingpointsmemo.com
2	www.dailykos.com
3	atrios.blogspot.com
4	www.washingtonmonthly.com
5	www.wonkette.com
6	instapundit.com
7	www.juancole.com
8	powerlineblog.com
9	americablog.blogspot.com
10	www.crooksandliars.com

Table 1. The Feeds That Matter for Politics Shows the Top Political Blogs Ranked Using Readership-Based Influence Metrics.

Modeling Influence in the Blogosphere

The blogosphere provides an interesting opportunity to study online social interactions including spread of information, opinion formation, and influence. Through original content and sometimes through commentary on topics of current interest, bloggers influence each other and their audience. We are working to study and characterize these social interactions by modeling the blogosphere and providing novel algorithms for analyzing social media content. Figure 2 shows a hypothetical blog graph and its corresponding flow of information in the *influence graph.*

Studies on influence in social networks and collaboration graphs have typically focused on the task of identifying key individuals who play an important role in propagating information. This is similar to finding authoritative pages on the web. Epidemic-based models like linear threshold and cascade models (Kempe, Kleinberg, and Tardos 2003 and 2005; Leskovec et al. 2007) have been used to find a small set of individuals who are most influential in a social network. However, influence on the web is often a function of topic. For example, Engadget's (engadget.com) influence is in the domain of consumer electronics, and Daily Kos's (dailykos.com) is in politics. A post in the former is unlikely to be very effective in influencing opinions on political issues even though Engadget is one of the most popular blogs on the web.

The other related dimension of influence is readership. With the large number of niches existing in the blogosphere, a blog that is relatively low ranked can be highly influential in this small community of interest. In addition, influence can be subjective and based on the interest of the users. Thus by analyzing the readership of a blog we gain insights into the community that is likely to be influenced by it.

We have implemented a system called Feeds That Matter (ftm.umbc.edu) (Java et al. 2007b) that aggregates subscription information across thousands of Bloglines users to automatically categorize blogs into different topics. Bloglines (bloglines.com) is a popular *feed reader* service that lets users manage subscriptions and monitor a number of feeds for any unread posts. Bloglines provides a feature that allows users to share their subscriptions. We conduct a study of the publicly listed OPML[1] feeds from 83,204 users consisting of a total of 2,786,687 subscriptions of which 496,879 are unique. A Blogline user's feeds are typically organized into named folders, such as *Podcasts* or *Politics*, and the folder structure is maintained in the OPML representation. The folder names can be used as an approximation of the topic that a user associated with a feed. By clustering related folders, we can induce an intuitive set of topics for feeds and blogs. Figure 3 shows a tag cloud of popular topics aggregated from readership information. Finally, we rank the feeds relevant to each of the topics generated. In our approach, we say that a feed is topically relevant and authoritative if many users have categorized it under similar folder names. For example, table 1 shows the top political blogs ranked using readership-based influence metrics.

An important component in understanding influence is to detect the sentiment and opinions expressed in blog posts. An aggregated opinion over

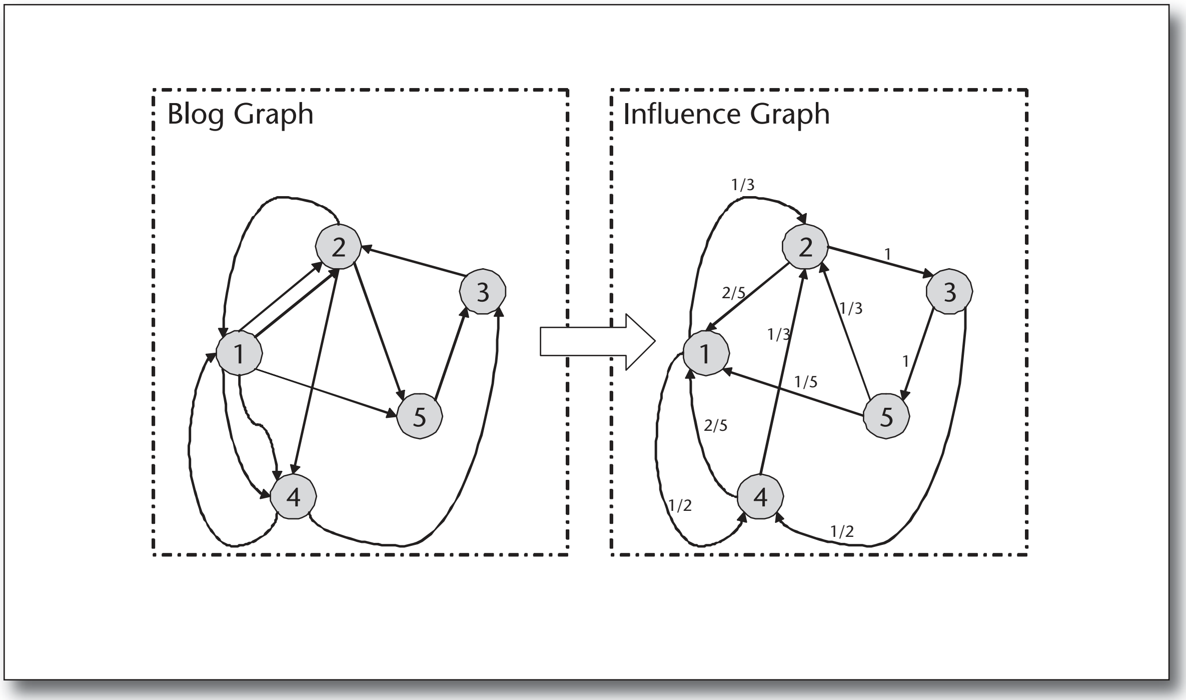


Figure 2. A Graph of Blogs and Web Links between Them Can Be Converted into an Influence Graph.

A link from *u* to *v* indicates that *u* is influenced by *v*. The edges in the influence graph are reverse of the blog graph to indicate this influence. Multiple edges indicate stronger influence and are weighed higher.

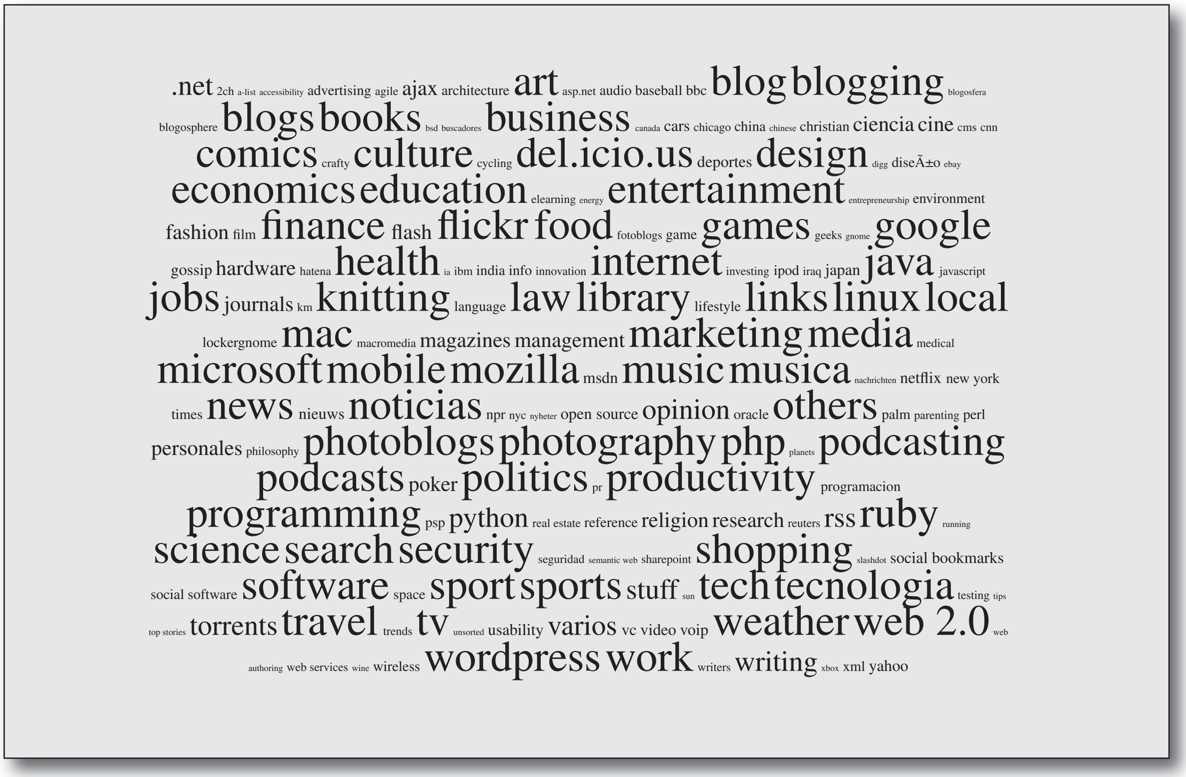


Figure 3. The Tag Cloud Generated from the Top 200 Folders before and after Merging Related Folders.

The size of the word is scaled to indicate how many users use the folder name.

many users is a predictor for an interesting trend in a community. Sufficient adoption of this trend could lead to a "tipping point" and consequently influence the rest of the community. The BlogVox system (Java et al. 2007a, Martineau et al. 2007, Balijepalli 2007) retrieves opinionated blog posts specified by ad hoc queries identifying an entity or topic of interest (for example, *March of the Penguins*). After retrieving posts relevant to a topic query, the system processes them to produce a set of independent features estimating the likelihood that a post expresses an opinion about the topic. These are combined using a Support Vector Machine (SVM)–based system and integrated with the relevancy score to rank the results.

Since blog posts are often informally written, poorly structured, rife with spelling and grammatical errors, and feature nontraditional content, they are difficult to process with standard language-analysis tools. Performing linguistic analysis on blogs is plagued by two additional problems: (1) the presence of spam blogs and spam comments and (2) extraneous noncontent including blog rolls, link rolls, advertisements, and sidebars. In the next section we describe techniques designed to eliminate spam content from a blog index. This is a vital task before any useful analytics can be supported on social media content.

In the following sections we also introduce a technique we call link polarity. We represent each edge in the influence graph with a vector of topic and corresponding weights indicating either positive or negative sentiment associated with the link for a web resource. Thus if a blog A links to a blog B with a negative sentiment for a topic T, influencing B would have little effect on A. Opinions are also manifested as biases. A community of iPod fanatics, for example, needs little or no convincing that it is a good product. Thus, attempting to influence an opinion leader in such already positively biased communities will have less impact. Using link polarity and trust propagation, we have demonstrated how like-minded blogs can be discovered and the potential of using this technique for more generic problems such as detecting trustworthy nodes in web graphs (Kale et al. 2007).

Existing models of influence have considered a static view of the network. The blogosphere, on the other hand, is extremely dynamic and "buzzy." New topics emerge and blogs constantly rise and fall in popularity. By considering influence as a temporal phenomenon, we can find key individuals that are early adopters or "buzz generators" for a topic. We propose an abstract model of the blogosphere that provides a systematic approach to modeling the evolution of the link structure and communities. Thus in order to model influence on the blogosphere, we need to consider topic, readership, community structure, sentiment, and time.

In the following sections, we provide a detailed description of various issues that need to be handled in order to model influence. Detecting influence and understanding its role in how people perceive and adopt a product or service provides a powerful tool for marketing, advertising, and business intelligence. This requires new algorithms that build on social network analysis, community detection, and opinion extraction.

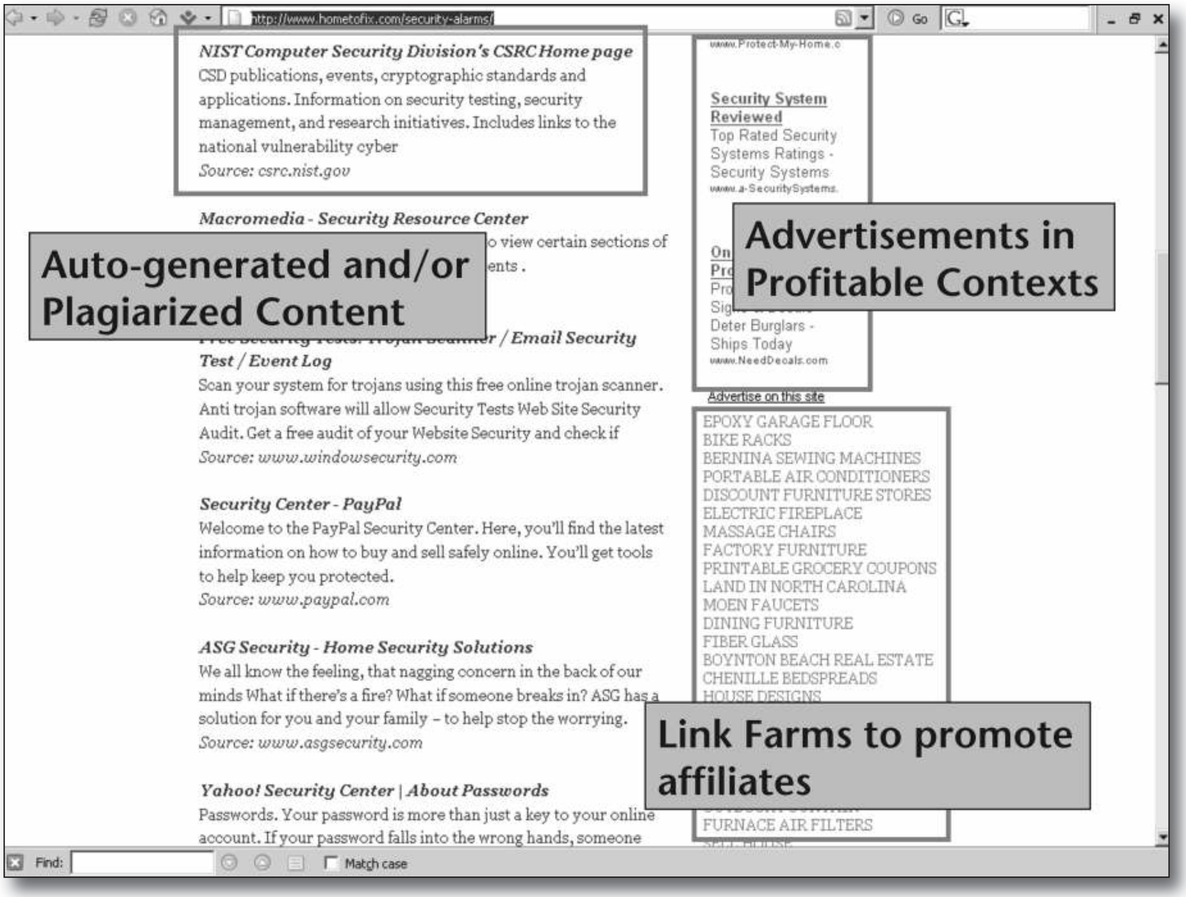
Detecting Blog Spam

As with other forms of communication, spam has become a serious problem in blogs and social media, both for users and for systems that harvest, index, and analyze generated content. Two forms of spam are common in blogs: spam blogs (also known as splogs), where the entire blog and hosted posts are machine generated, and spam comments, where authentic posts feature machine-generated comments. Though splogs continue to be a problem for web search engines and are considered a special case of web spam, they present a new set of challenges for blog analytics. Given the context of this paper and the intricacies of indexing blogs (Mishne 2007), we limit our discussion to splogs.

Blog search engines index new blog posts by processing pings from update ping servers, intermediary systems that aggregate notifications from updated blogs. Scores of spam pages infiltrate at these ping servers, increasing computational requirements, corrupting results, and eventually reducing user satisfaction. We estimate that more than 50 percent of all pings are from spam sources (Kolari, Java, and Finin 2006). Two kinds of spam content sources are prevalent:

Nonblogs are pages that attempt to increase the visibility of hosted and linked-to content by feigning to be blogs to leverage higher trust and quicker indexing by web search engines. Though such nonblogs need not necessarily be spam from the web perspective, they do constitute spam in the context of the blogosphere. An example of one such nonblog is an Amazon affiliate (third-party vendor of products) book-selling site that pings an update ping server.

Spam blogs constitute the second kind of spam. These blogs, created using splog creation tools[2], are either fully or partly machine generated. Splogs have two, often overlapping, motives. The first is the creation of blogs containing gibberish or hijacked content from other blogs and news sources with the sole purpose of hosting profitable context-based advertisements. The second is the creation of blogs that realize link farms intended to increase the ranking of affiliate sites (blogs or nonblog web pages). One such splog is shown in figure 4.



http://www.hometofix.com/security-alarms/

NIST Computer Security Division's CSRC Home page
CSD publications, events, cryptographic standards and applications. Information on security testing, security management, and research initiatives. Includes links to the national vulnerability cyber
Source: csrc.nist.gov

Macromedia - Security Resource Center
o view certain sections of
ents .

Free Security Tests: Trojan Scanner / Email Security Test / Event Log
Scan your system for trojans using this free online trojan scanner. Anti trojan software will allow Security Tests Web Site Security Audit. Get a free audit of your Website Security and check if
Source: www.windowsecurity.com

Security Center - PayPal
Welcome to the PayPal Security Center. Here, you'll find the latest information on how to buy and sell safely online. You'll get tools to help keep you protected.
Source: www.paypal.com

ASG Security - Home Security Solutions
We all know the feeling, that nagging concern in the back of our minds What if there's a fire? What if someone breaks in? ASG has a solution for you and your family – to help stop the worrying.
Source: www.asgsecurity.com

Yahoo! Security Center | About Passwords
Passwords. Your password is more than just a key to your online account. If your password falls into the wrong hands, someone

www.Protect-My-Home.c

Security System Reviewed
Top Rated Security Systems Ratings - Security Systems
www.a-SecuritySystems.

On
Pr
Pro
Sig
Deter Burglars - Ships Today
www.NeedDecals.com

Advertise on this site

EPOXY GARAGE FLOOR
BIKE RACKS
BERNINA SEWING MACHINES
PORTABLE AIR CONDITIONERS
DISCOUNT FURNITURE STORES
ELECTRIC FIREPLACE
MASSAGE CHAIRS
FACTORY FURNITURE
PRINTABLE GROCERY COUPONS
LAND IN NORTH CAROLINA
MOEN FAUCETS
DINING FURNITURE
FIBER GLASS
BOYNTON BEACH REAL ESTATE
CHENILLE BEDSPREADS
HOUSE DESIGNS
FURNACE AIR FILTERS

Go
Find:
Match case
Auto-generated and/or Plagiarized Content
Advertisements in Profitable Contexts
Link Farms to promote affiliates

Figure 4. An Example Splog.

This example of a splog contains plagiarized content, promotes other spam pages by linking to them, and hosts high-paying advertisements automatically selected to match the splog's content.

Detecting Splogs

Over the past year (Kolari 2007) we have developed techniques to detect spam blogs as they fit the overall architecture (figure 5), arrived at through our discussions with practitioners. Our existing and continuing work has explored all aspects of this architecture. We discuss highlights of our effort based on splog detection using blog home pages with local and relational features. Interested readers are referred to Kolari et al. (2006) and Kolari, Finin, and Joshi (2006) for further details.

Results reported in the rest of this section are based on a seed data set of 700 positive (splogs) and 700 negative (authentic blog) labeled examples containing the entire HTML content of each blog home page. All of the models are based on Support Vector Machines or SVMs (Boser, Guyon, and Vapnik 1992), which are known to perform well in classification tasks (Boser, Guyon, and Vapnik 1992). We use a linear kernel with the top features chosen using mutual information, and models are evaluated using leave-one-out cross-validation. We view detection techniques as local and relational, based on feature types used.

Local Features

A blog's local features can be quite effective for splog detection. A *local feature* is one that is completely determined by the contents of a single web page, that is, it does not require following links or consulting other data sources. A local model built using only these features can provide a quick assessment of the authenticity of blogs. We have experimented with many such models, and our results are summarized in figure 6.

Words. To verify their utility, we created a bag-of-words for the samples based on their textual content. We also analyzed discriminating features by ordering features based on weights assigned to them by the linear kernel. It turns out that the model was built around features that the human eye would have typically overlooked. Blogs often contain content that expresses personal opinions, so words like *I*, *we*, *my*, *what* appear commonly on authentic blog posts. To this effect, the bag-of-words model is built on an interesting "blog content genre." In general, such a content genre is not seen on the web, which partly explains why spam

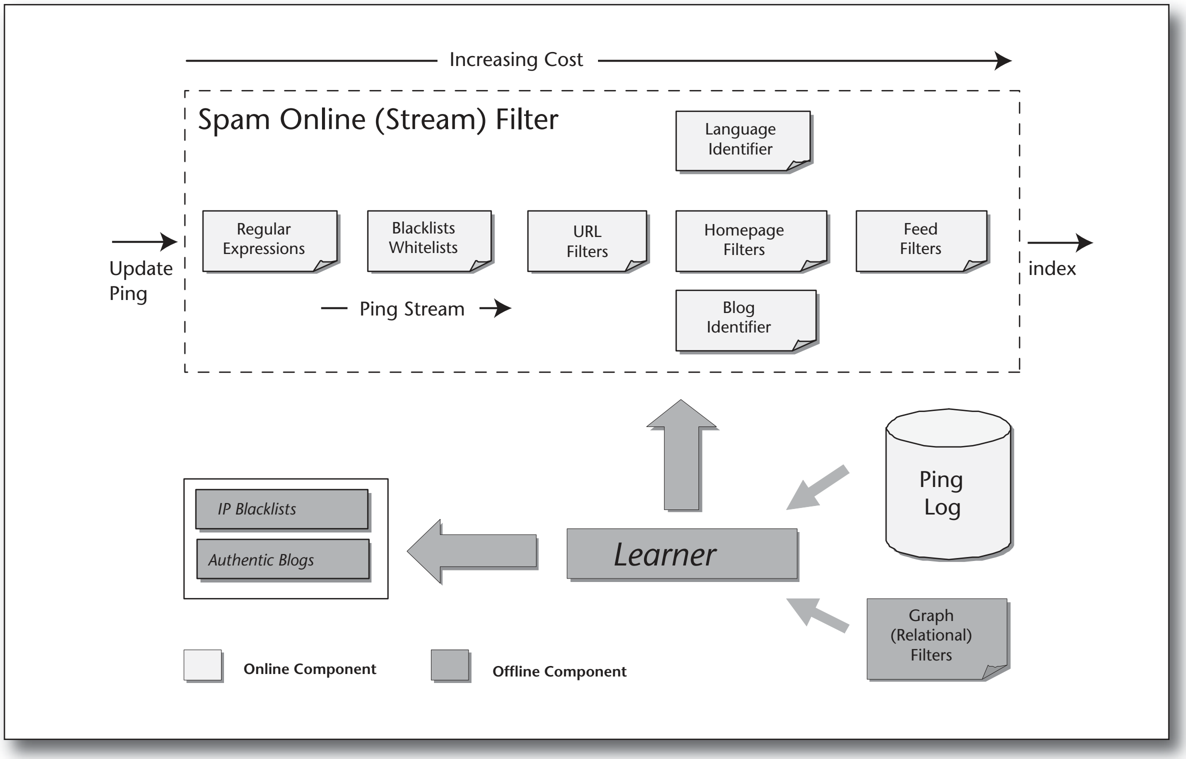


Figure 5. Our Online Spam Detection System.

In our online spam detection system, pings from existing ping servers are aggregated and pass through a stream-based filtering system. Subfilters are ordered based on their cost of filtering with each deciding whether to reject a ping as spam or pass it through. Relational techniques are used in the offline component. The system uses all these detection techniques together to adapt and coevolve through a learning component.

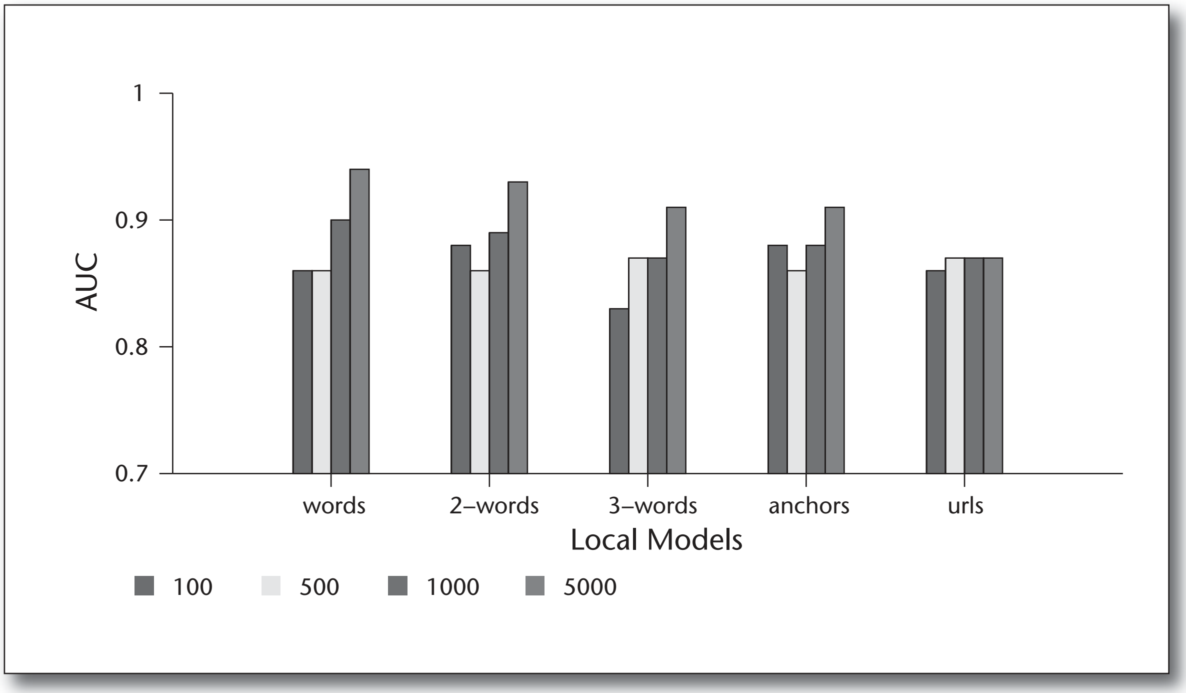


Figure 6. The Performance of Local Models, as Measured by the Standard, Area under the Curve Metric, Varies for Different Feature Types and Sizes.

detection using local textual content is less effective there.

Word N-Grams. An alternative methodology to using textual content for classification is the bag-of-word-N-Grams, where *N* adjacent words are used as a feature. We evaluated both bag-of-word-2-Grams and bag-of-word-3-Grams, which turned out to be almost as effective as bag-of-words. Interesting discriminative features were observed in this experiment. For instance, text like "comments-off" (comments are usually turned off in splogs), "new-york" (a high-paying advertising term), "in-uncategorized" (spammers do not bother to specify categories for blog posts) are features common to splogs, whereas text like "2-comments," "1-comment," "i-have," "to-my" were some features common to authentic blogs. Similar features ranked highly in the 3-word gram model.

Tokenized Anchors. Anchor text is the text that appears in an HTML link (that is, between the <a…> and </a> tags.) and is a common link-spamming technique around profitable contexts. We used a bag-of-anchors feature, where anchor text on a page, with multiple word anchors split into individual words, is used. Note that anchor text is frequently used for web page classification, but typically to classify the target page rather than the one hosting the link. We observed that "comment" and "flickr" were among the highly ranked features for authentic blogs.

Tokenized URLs. Intuitively, both local and outgoing URLs can be used as effective attributes for splog detection. This is motivated by the fact that many URL tokens in splogs are in profitable contexts. We term these features as bag-of-urls, arrived at by tokenizing URLs using "/" and "..". Results indicate this can be a useful approach complementing other techniques.

Relational Features

A global model is one that uses some nonlocal features, that is, features requiring data beyond the content of the web page under test. We have investigated the use of link distributions to see if splogs can be identified once they place themselves on the blog (web) hyperlink graph. The intuition is that authentic blogs are very unlikely to link to splogs and that splogs frequently do link to other splogs. We have evaluated this approach by extending our seed data set with labeled in-links and out-links, to achieve AUC (area under curve) values of close to 0.85. Interested readers are referred to (Kolari et al. 2006, Kolari 2007) for further details.

Future Challenges

Though current techniques work well, the problem of spam detection is an adversarial challenge. In our continuing efforts we are working toward better addressing concept drift and leveraging community and relational features. The problem of spam in social media is now extending well beyond blogs and is quite common in popular social tools like Myspace and Facebook. The nature of these social tools demands additional emphasis on relational techniques, a direction we are exploring as well.

Recognizing Blogosphere Communities

Underlying most forms of social media is the concept of a community of people. Identifying these communities, and their possible subcommunities, continues to be a ubiquitous and important task. Most work on community detection (Gibson, Kleinberg, and Raghavan 1998) is based on the analysis of the networks associated with social system.

We have addressed the problem of community detection in the blogosphere by modeling trust and influence (Kale et al. 2007, Kale 2007). Our approach uses the link structure of a blog graph to associate sentiments with the links connecting blogs. Such links are manifested as a URL that blogger *A* uses in his blog post to refer to blogger *B*'s post. We call this sentiment *link polarity*, and the sign and magnitude of this value is based on the sentiment of text surrounding the link. These polar edges are evidence of bias, trust, or distrust between respective blogs. We then use trust propagation models to "spread" the polarity values from a subset of nodes to all possible pairs of nodes. We have evaluated this technique of using trust propagation on polar links in the domain of political blogs by predicting the "like-mindedness" of blogs oriented toward either a Democratic or Republican position. In order to determine a blog's bias, we compute its trust/distrust score from a seed set of influential blogs (discussed later) and use a hand-labeled data set to validate our results. More generally, we address the problem of detecting all such nodes that a given node would trust even if not directly connected to them.

Link Polarity

The term *link polarity* represents the opinion of the source blog about the destination blog. In order to determine the sentiment based on links, we analyze the section of text around the link in the source blog post to determine the sentiment of the source blogger about the destination blogger. The text neighboring the link provides direct meaningful insight into blogger A's opinion about blogger B. Hence, we consider a window of *X* characters (*X* is a variable parameter for our experimental validations) before and after the link. Note that this set of 2*X* characters does not include HTML tags.

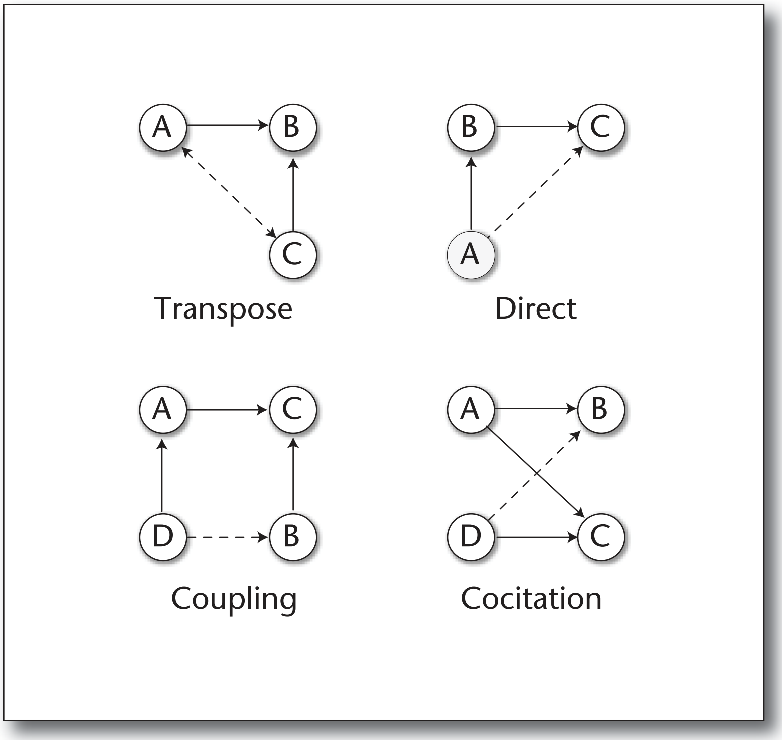


Figure 7. Four Graphs Representing Our Atomic Propagation Patterns.

The solid and dotted arrows represent known and inferred trust scores, respectively. The first, for example, indicates that if A and C trust B, then A and C are likely to trust each other.

For our requirements, we do not need to employ complex natural language-processing techniques since bloggers typically convey their bias about the post/blog pointed to by the link in a straightforward manner. Hence, we use a manually created lexicon of positively and negatively oriented words and match the token words from the set of 2*X* characters against this corpus to determine the polarity. Since bloggers frequently use negation of sentimental words to indicate bias about another blog post ("What B says is not bad"), our corpus includes simple bi-gram patterns of the form "not positive/negative word."

We adopted the following formula for calculating the link polarity between two posts:

Polarity = $(N_p - N_n) / (N_p + N_n)$

where N_p is the number of positively oriented words and N_n is the number of negatively oriented words. Notice that our formula incorporates zero polarity links automatically. The term in the denominator ensures that the polarity is weighed according to the number of words matched against the lexicon. We use summation as the aggregation technique for computing the polarity between two blogs. For our experiments, we choose a domain with a low probability of "off-the-topic" posts within a single blog, hence the notion of summing post-post polarity values to yield a blog-blog polarity value holds.

Trust Propagation

Since blog graphs are not always densely connected, we will not have the trust scores between many pairs of nodes. Hence, we have investigated techniques for inferring a trust relationship for nodes where one is not explicitly known. Guha and colleagues (2004) have used a framework to spread trust in a network bootstrapped by a known set of trusted nodes. Their approach uses a "belief matrix" to represent the initial set of beliefs in the graph. This matrix is generated through a combination of known trust and distrust among a subset of nodes. This matrix is then iteratively modified using "atomic propagations." The "atomic propagation" step incorporates direct propagation, cocitation, and transpose trust and trust coupling as described in figure 7. Finally, a "rounding" technique is applied on the final matrix to produce absolute values of trust between all pair of nodes.

In order to form clusters after the step of trust propagation, we take the approach of averaging trust scores for all blog nodes from a predefined set of "influential" nodes belonging to each community. A positive trust score indicates that the blog node belongs to the community influenced by the trusted node of that community. Specifically, we selected the top three influential (using the number of in-links as the measure) Democratic and Republican bloggers. A positive trust score for a blog from the top three Democratic blogs indicates that it belongs to the Democratic cluster and a negative score indicates that it is a Republican blogger.

Experiments

We choose political blogs as our domain; one of the major goals of the experiments was to validate that our proposed approach can correctly classify the blogs into two sets: Republican and Democratic. Through some manual analysis of the political blogs, we observed that the link density among political blogs is reasonably high and hence we could deduce the effectiveness of our approach by running our algorithms over a fairly small number of blogs.

Guha and colleagues argue that "one step distrust" provides the best trust propagation results in their domain of experiments. They propose the notion of "trust and distrust" between two nodes in the graph where the same set of two nodes can trust or distrust each other. The "one step distrust" approach uses the "trust matrix" as the belief matrix. However, we believe that in our domain the initial belief matrix should incorporate both trust and distrust (positive and negative polarities from blog A to blog B). Hence, we use the difference between the trust and distrust matrices as our initial

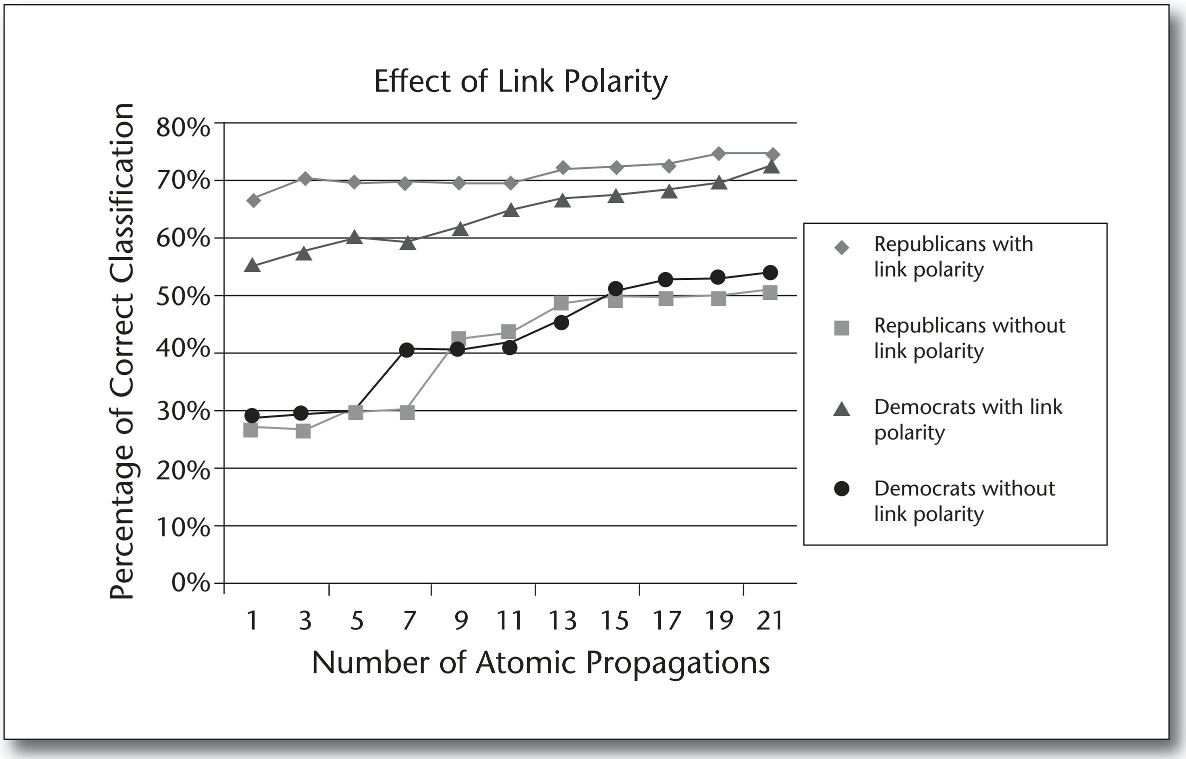


Figure 8. Our Experiments Show That Using Polar Links for Classification Yields Better Results Than Plain Link Structure.

belief matrix. We experimented with various values of the "alpha vector" (the vector used to define the fractional weights for atomic propagation parameters) to confirm Guha and colleagues' conclusion that using the values they proposed {0.4, 0.4, 0.1, 0.1} yields best results. Further, Guha et al. recommend performing "atomic propagations" approximately 20 times to get best results; we took the approach of iteratively applying atomic propagations till convergence, and our experiments indeed indicate a value close to 20.

Test Data Set. Our test data set consists of a blog graph created from the link structure of Buzzmetrics' data set.[3] The data set consists of about 14 million weblog posts from three million weblogs collected by Nielsen BuzzMetrics for May 2006. The data is annotated with 1.7 million blog-blog links from the Buzzmetrics data set.

Reference Data Set. Adamic and Glance (2005) provided us with a reference data set of 1490 blogs with a label of *Democratic* and *Republican* for each blog. Their data on political leaning is based on analysis of blog directories and manual labeling and has a timeframe of the 2004 presidential elections.

Our test data set from Buzzmetrics did not provide a classified set of political blogs. Hence, for our experiments we used a snapshot of Buzzmetrics that had a complete overlap with our reference data set to validate the classification results. The snapshot contained 297 blogs, 1309 blog-blog links, and 7052 post-post links. The reference data set labeled 132 blogs as Republicans and 165 blogs as Democrats (there did not exist any *neutral* labels).

Effect of Link Polarity. The results in figure 8 indicate a clear improvement on classifying Republican and Democratic blogs by applying polar weights to links followed by trust propagation. We get a "cold-start" for Democratic blogs, and we observe that the overall results are better for Republican blogs than Democratic blogs. The results being better for Republican blogs can be attributed to the observations from Adamic and Glance (2005) that Republican blogs typically have a higher connectivity than Democratic blogs in the political blogosphere.

We believe that the idea of *polar links* is quite useful and can be applied to multiple domains. The main contribution of this work lies in applying trust propagation models over polar links. We demonstrated one such application in the domain of the political blogosphere where we used natural language processing to deduce the link polarity. We would like to emphasize that the specific techniques to generate polar links is orthogonal to our main contribution, and our approach can be easily adapted to different domains for modeling trust and detecting predefined communities.

Property	ER Model	BA Model	Blogosphere	Our Simulation
Type	undirected	undirected	directed	directed
Degree distribution	poisson	power law	power law	power law
Slope [inlinks, outlinks]	N/A	[2.08, –]	[1.66–1.8, 1.6–1.75]	[1.7–2.1, 1.5–1.6]
Average degree	constant (for given *p*)	constant (adds *m* edges)	increases	increases
Component distribution	N/A (undirected)	N/A (undirected)	Power law	Power law
Correlation coefficient	–	1 (fully preferential)	0.024 (WWE)	0.1
Average clustering coefficient	0.00017	0.00018	0.0235 (WWE)	0.0242
Reciprocity	N/A (undirected)	N/A (undirected)	0.6 (WWE)	0.6

Table 2. Various Graph Properties for Two Popular Network Models (ER and BA), an Empirical Study of a Blogosphere Sample, and Our Simulated Model.

A Generative Model for the Blogosphere

The blog analysis techniques described in the earlier sections are data intensive. They require large amounts of blog data that must be obtained by crawling the blogosphere. Moreover, if the collection is not comprehensive, attention needs to be paid to ensure a representative sample. Once collected, the data must be cleaned to remove spurious spam blogs, or splogs (Kolari et al. 2006) and preprocessed in various ways (Leskovec et al. 2007; Shi, Tseng, and Adamic 2007). To overcome the similar difficulty in web analysis, various graph models have been proposed for the structural and statistical analysis, including the Barabasi model (Barabasi and Albert 1999) and Pennock model (Pennock et al. 2002). However, these models are not suitable for generating the blog graphs. While the blog networks resemble many properties of web graphs, the dynamic nature of the blogosphere and the evolution of the link structure due to blog readership and social interactions are not well expressed by the existing models.

There are several motivations for developing a generative model for the blogosphere. First, we hope that such a model might help us understand various aspects of the blogosphere at an abstract level. Secondly, noticing that a portion of the blog graph deviates from the blogosphere graph can signal that something is amiss. Spam blogs, for example, often form communities whose structural properties are very unlike those of naturally occurring blogs. Third, a generative model can be used to create artificial data sets of varying sizes that simulate portions of the blogosphere, which is often useful for testing and comparing algorithms and systems. For example, testing a model with hidden variables that measures a blog's influence can benefit from the simulated blogosphere with different blog graph structures.

Modeling Blogger Characteristics

Our generative model to construct blog graphs is based on the general characteristics of the bloggers as observed in a recent PEW Internet Survey[4], which can be summarized as follows: (1) Blog writers are enthusiastic blog readers. (2) Most bloggers post infrequently. (3) Blog readership can be inferred through blog rolls, a list of links to related or friends' blogs. Active bloggers are more likely to have a blog roll and follow it regularly.

Our model uses the elements of the existing preferential attachment (Barabasi and Albert 1999) and random attachment models (Chung and Lu 2006). Each blogger is assumed to be reading, writing, or being idle according to the preferential selection of the bloggers. This helps to capture the linking pattern arising in the blogosphere through local interactions. *Local interactions* refer to the interaction of the bloggers among the other blogs that are generally connected to them either by an in-link or out-link. We have studied the properties including the degree distributions, degree correlation, clustering coefficient, average degree, reciprocity, and the distribution of connected components. To the best of our knowledge, there exist no general models to generate the blog and post networks that possess the properties observed in the real-world blogs. Table 2 gives a quick comparison of the properties of the existing web models and shows the need for a model for the blogosphere.

Defining Blog and Post Networks

Our model of the blogosphere includes two related

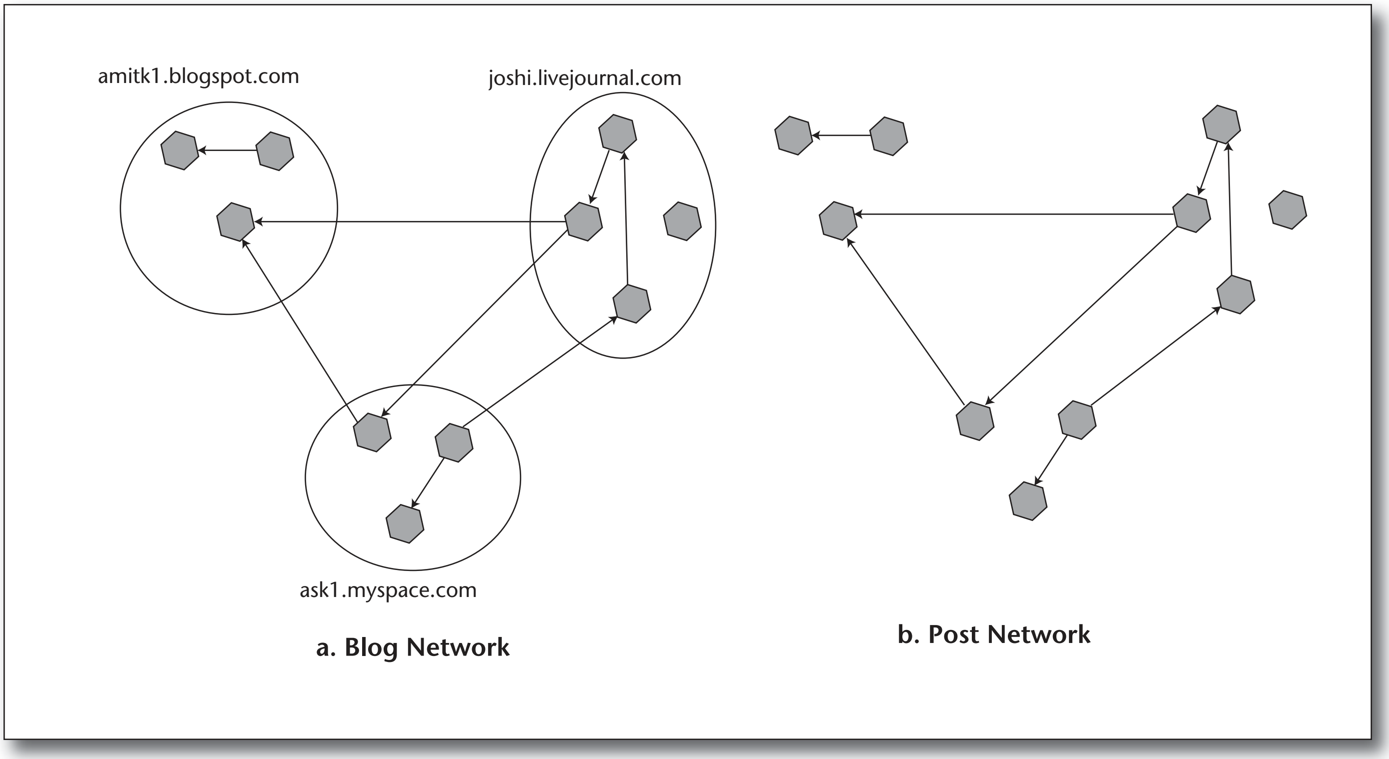


Figure 9. The Graph Representation for the Blogosphere Includes Both a Blog Network and Post Network.

networks, one for blogs and one for their posts. The *blog network* (Figure 9a) is defined as a network of blogs obtained by collapsing all directed post links between blog posts into directed edges between blogs. Blog networks give a macroscopic view of the blogosphere and help to infer a social network structure, under the assumption that blogs that are "friends" link each other more often. The *post network* (Leskovec et al. 2007) shown in figure 9b is formed by ignoring the posts' parent blogs and focusing on the link structure among posts only. Each post also has a time stamp of the post associated with it. Post networks give a microscopic view of the blogosphere with details such as which post linked to which other post and at what time.

Design Considerations

In designing our model we relied on our experience in analyzing, using, and generating blog content as well as an investigation into previous models for social media graphs. We describe how we addressed some of these observations in our model in the following paragraphs.

Linking of New Blogs. The new blogger (blog node)[5] may join the existing network by linking to a popular blog (a blog node with high in-degree) or may not link at all. Many web models using continuous growth (Chung and Lu 2006) also use similar techniques for addition of the new node.

Linking in Blogosphere. Generally, active bloggers read several posts and tend to link to some of the posts that they read recently, but the only "observable behavior" is the creation of a link to the read post (destination). We model this behavior by having the blogger keep track of the recently read blog posts and link to them.

Linking to a Post. Leskovec et al. (2007) observed that any post gathers most of its in-links within 24 hours of post time. We approximate this behavior by linking to recent posts (within a fixed window) of the visited blog when our blogger visits any blog node.

Blogger Neighborhood. Active bloggers tend to subscribe to the well-known blogs of interest and read the subscriptions regularly forming blog readership (Leskovec et al. 2007). Hence we see that the blogger interactions are largely concentrated in the blog neighborhood, that is, nodes connected by either in-links or out-links.

Use of Emerging Tools in the Blogosphere. New tools in the blogosphere help to discover popular blogs. They include, for example, blog search engines such as Technorati, social bookmarking and ranking systems like Del.icio.us and Digg, blog classification systems like Feeds That Matter, and trend discovery systems like BlogPulse. The availability and use of these systems mean that blog post reads are not totally random but biased to-

Blog Network	ICWSM	WWE	Simulation
Total Blogs	159,036	650,660	650,000
Blog-blog Links	435,675	1,893,187	1,451,069
Unique links	245,840	648,566	1,158,803
Average Degree	5.47	5.73	4.47
n-degree Distribution	–2.07	–2.0	–1.71
ut-degree Distribution	–1.51	–1.6	–1.76
Degree Correlation	0.056	0.002	0.10
Diameter	14	12	6
Largest WCC Size	96,806	263,515	617,044
Largest SCC Size	4,787	4,614	72,303
Clustering Coefficients	0.04429	0.0235	0.0242
Percent Reciprocity	3.03	0.6838	0.6902

Table 3. Comparing a Simulated Blog Network with Two Data Sets Based on B logs Harvested from the Web Using a Number of Standard Graph Metrics.

Post Network	ICWSM	WWE	Simulation
Total posts	1,035,361	1,527,348	1,380,341
Post-post links	1,354,610	1,863,979	1,451,069
Unique links	458,950	1,195,072	1,442,525
Avg post out-links	1.30	1.22	1.051
Average degree	2.62	2.44	2.10
n-degree distribution	–1.26	–2.6	–2.54
ut-degree distribution	–1.03	–2.04	–2.04
Degree correlation	–0.113	–0.035	–0.006
Diameter	20	24	12
Largest WCC size	134,883	262,919	1,068,755
Largest SCC size	14	13	3
Clustering coefficients	0.0026	0.00135	0.00011
Percent Reciprocity	0.029	0.021	0.01

Table 4. Comparing Properties of the Post Network Extracted from the ICWSM and WWE Data Sets and the Simulated Blogosphere Data.

wards the popularity of the blogs that are initially not known to the blogger.

Conversations through Comments and Trackbacks. The exchange of links among bloggers through comments and trackbacks leads to *higher reciprocity* (that is, through reciprocal links) in the blogosphere than the random networks. Bloggers tend to link to the blogs to which they have linked in the past either through comments, trackbacks, or general readership. We expect these local interactions to provide for a *higher clustering coefficient* (as observed in the blogosphere) than the random networks.

Activity in the Blogosphere. Not all bloggers are "active" (either reading or writing) at all times. Only a small portion of the blogosphere is *active* with the remainder identified as *idle*. This activity can be approximated by observing the number of links created within a time span. We use a *super linear growth function* to model the activity as defined by Leskovec. The out-links from a blog can be considered as the measure of an *active blog writer* because an active writer will naturally look for more interesting sources to link to. The reverse may not be true that the blogger who reads a lot also writes more.

Experiments and Results

The preferential attachment model produces the power law degree distributions in an *undirected net-*

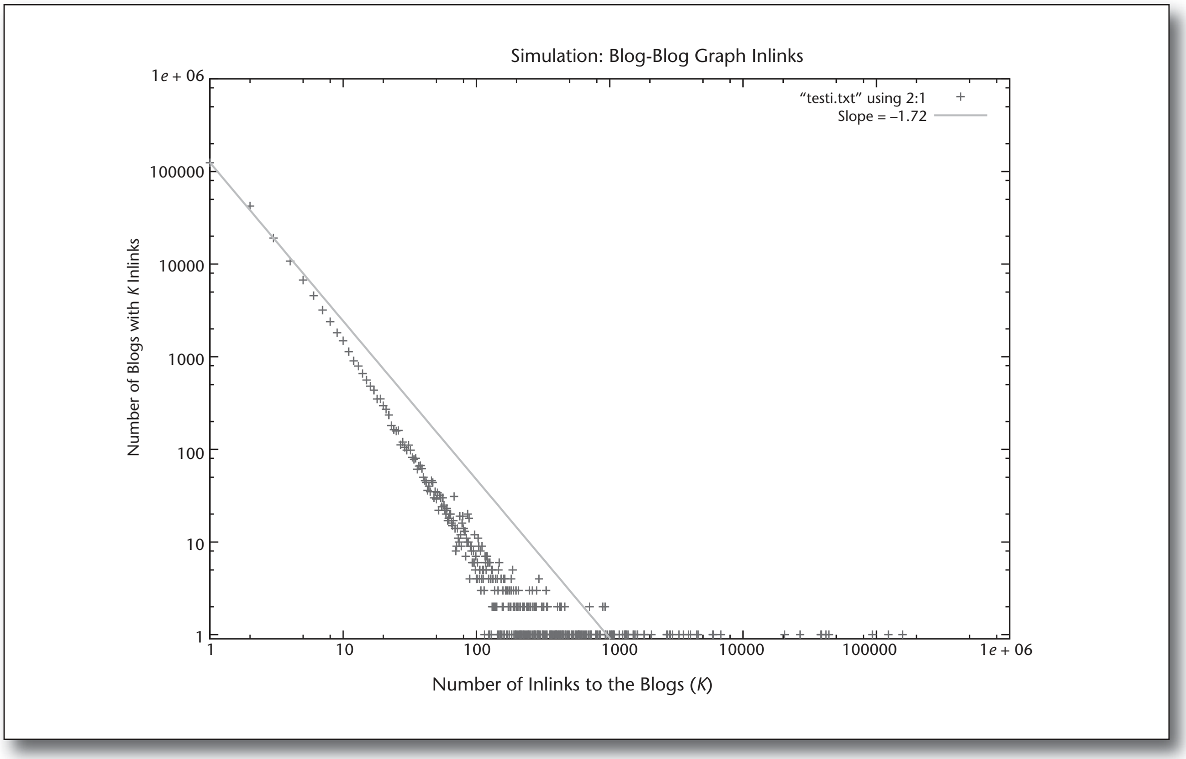


Figure 10. Simulation: Blog In-Links Distribution.

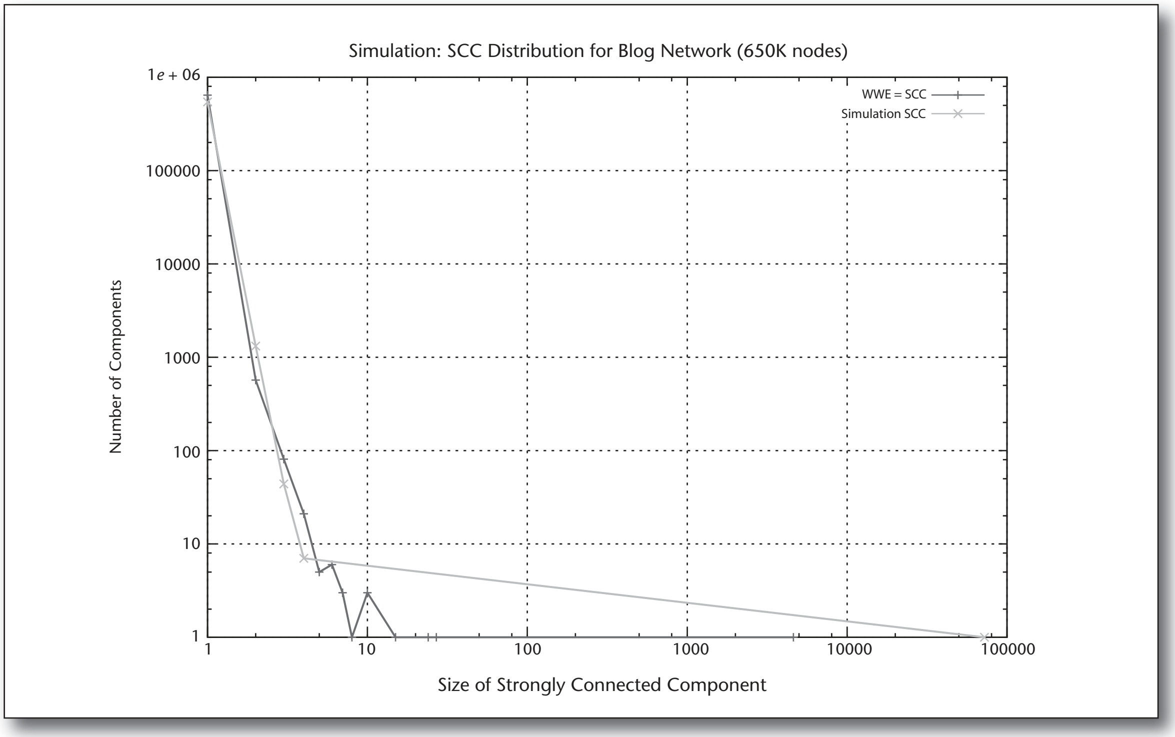


Figure 11. Distribution of SCC in Blog Network (Simulation and Blogosphere).

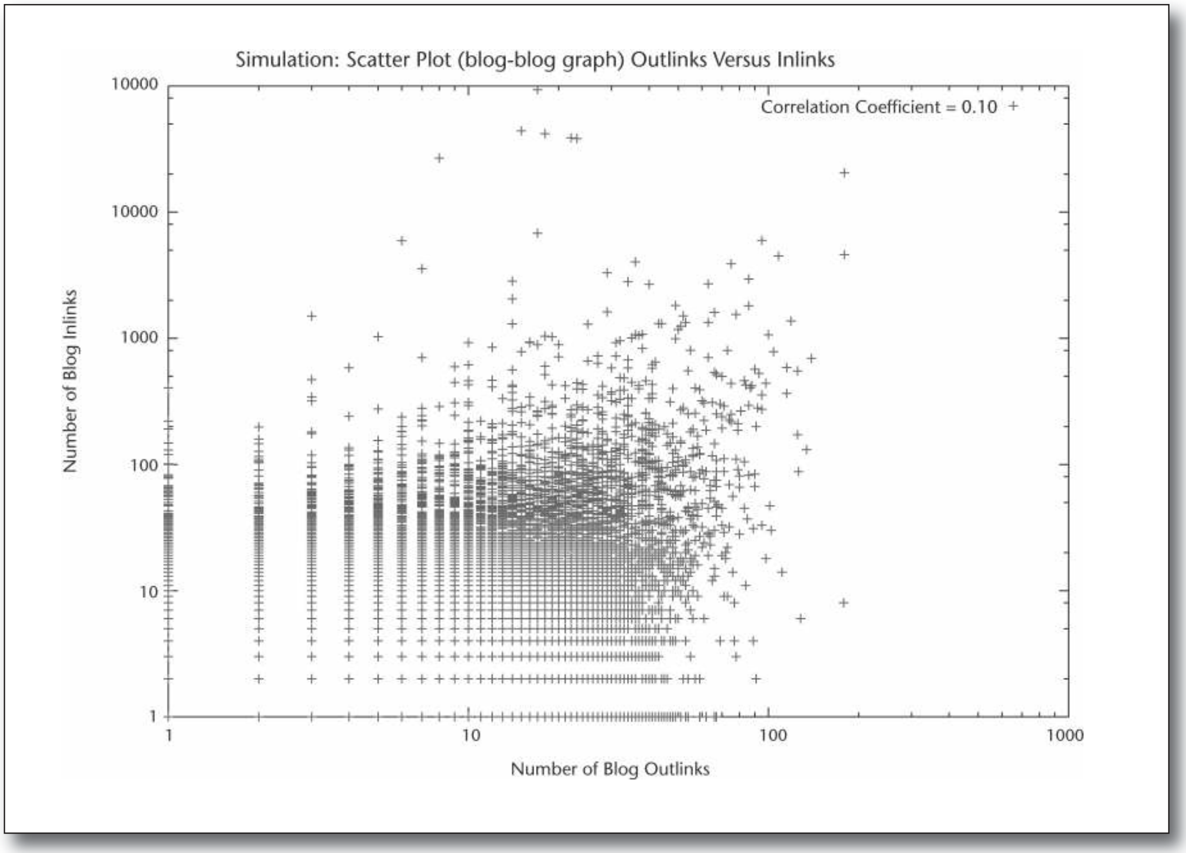


Figure 12. A Scatter Plot Showing the Out-Degree and In-Degree of the Simulated Blog Network.

work. However, this model is not defined for a *directed network*. The model proposed by Pennock captures the random behavior but does not capture the local interactions among nodes in the graph. We use the "alpha preferential attachment" model proposed by Chung to obtain power law degree distributions in a directed graph. We have modified this model to reflect local interaction among the bloggers by using preferential attachment among neighboring nodes. The details of our algorithm can be found in Karandikar (2007).

Part of our evaluation is done by comparing the distinguishing properties of the real blog graphs with the results of our simulation. These properties were verified against two large blog data sets available for researchers: WWE, a data set developed for the Workshop on the Web Logging Ecosystem held at the 2006 Web Conference, and ICWSM, a data set provided by the 2007 International Conference on Web Logs and Social Media. See tables 3 and 4.

Figure 10 shows the power law curve for in-link distribution observed in the blog network of the "simulated" blogosphere. Similarly, figure 11 compares the distribution of the strongly connected components (SCC) in the simulation and the blogosphere.

Figure 12 shows the scatter plot for in-degree and out-degree correlations in the simulated blogosphere. The plot shows the low degree correlation as observed in the real blogosphere by Leskovec and colleagues.

Being able to generate synthetic blog data is useful for testing and evaluating algorithms for analyzing and extracting information from the blogosphere. The utility, however, depends on the generated graphs being similar to the actual ones in key properties. We have created a model that accounts for several key features that influence how the blogosphere grows and its resulting structure. By selecting appropriate parameters for our model, we can generate graphs that more closely approximate observed blogosphere network properties than previous network models.

Conclusion

Social media systems are increasingly important on the web and account for a significant fraction of new content. The various kinds of social media are alike in that they all have rich underlying network structures that provide metadata and context that can help when extracting information from their content. We have described some initial results from ongoing work that is focused on extracting, modeling, and exploiting this structural information from the underlying networks.

As the web continues to evolve, we expect that the ways people interact with it, as content consumers as well as content providers, will also change. The result, however, will continue to represent an interesting and extravagant mixture of underlying networks—networks of individuals, groups, documents, opinions, beliefs, advertisements, and scams. These interwoven networks present new opportunities and challenges for extracting information and knowledge from them.

Acknowledgements

We thank Dr. James Mayfield, Justin Martineau, and Sandeep Balijepalli for their contributions to the work on sentiment detection. Partial support for this research was provided by the National Science Foundation (awards ITR-IIS-0325172 and NSF-ITR-IDM-0219649) and IBM.

Notes

1. OPML, or outline processor markup language, is an XML format commonly used to share lists of web-feed URLs.

2. See Tim Finin's "Splog Software from Hell" (ebiquity.umbc.edu/blogger/splog-softwarefrom-hell).

3. www.nielsenbuzzmetrics.com.

4. See Bloggers: A Portrait of the Internet's New Storytellers by A. Lenhart and S. Fox (www.pewinternet.org/PPF/r/186/report_display.asp).

5. *Blog* and *blogger* are synonymous, and the exact meaning is evident from the context.

References

Adamic, L. A., and Glance, N. 2005. The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In *Proceedings of the Third International Workshop on Link Discovery* (LinkKDD '05), 36–43. New York: Association for Computing Machinery.

Balijepalli, S. 2007. Blogvox2: A Modular Domain Independent Sentiment Analysis System. Master's thesis, Department of Computer Science, University of Maryland, Baltimore County, Baltimore, MD.

Barabasi, A. L., and Albert, R. 1999. Emergence of Scaling in Random Networks. *Science* 286(5439): 509.

Boser, B. E.; Guyon, I. M.; and Vapnik, V. N. 1992. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (COLT '92), 144–152. New York: Association for Computing Machinery.

Chung, F., and Lu, L. 2006. *Complex Graphs and Networks* (CBMS Regional Conference Series in Mathematics). Boston: American Mathematical Society.

Gibson, D.; Kleinberg, J.; and Raghavan, P. 1998. Inferring Web Communities from Link Topology. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia* (HYPERTEXT '98), 225–234. New York: ACM Press.

Guha, R.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2004. Propagation of Trust and Distrust. In *Proceedings of the Thirteenth International Conference on World Wide Web* (WWW '04), 403–412. New York: Association for Computing Machinery.

Java, A.; Kolari, P.; Finin, T.; Joshi, A.; Martineau, J.; and Mayfield, J. 2007a. The BlogVox Opinion Retrieval System. In *Proceedings of the Fifteenth Text Retrieval Conference*. Washington, DC: National Institute of Standards and Technology.

Java, A.; Kolari, P.; Finin, T.; Joshi, A.; and Oates, T. 2007b. Feeds That Matter: A Study of Bloglines Subscriptions. In *Proceedings of the Second International Conference on Web Logs and Social Media*. Menlo Park, CA: AAAI Press.

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning* (ECML '98), 137–142. Berlin: Springer-Verlag.

Kale, A. 2007. Modeling Trust and Influence in Blogosphere Using Link Polarity. Master's thesis, Department of Computer Science, University of Maryland, Baltimore County, Baltimore, MD.

Kale, A.; Karandikar, A.; Kolari, P.; Java, A.; Joshi, A.; and Finin, T. 2007. Modeling Trust and Influence in the Blogosphere Using Link Polarity. In *Proceedings of the Second International Conference on Web Logs and Social Media* (Short Paper). Menlo Park, CA: AAAI Press.

Karandikar, A. 2007. Generative Model to Construct Blog and Post Networks in Blogosphere. Master's thesis, Department of Computer Science, University of Maryland, Baltimore County, Baltimore, MD.

Kempe, D.; Kleinberg, J. M.; and Tardos, E. 2003. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 137–146. New York: Association for Computing Machinery.

Kempe, D.; Kleinberg, J. M.; and Tardos, E. 2005. Influential Nodes in a Diffusion Model for Social Networks. In *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming, Lecture Notes in Computer Science* 3580, 1127–1138. Berlin: Springer.

Kolari, P. 2007. Detecting Spam Blogs: An Adaptive Online Approach. Ph.D. Dissertation, Department of Computer Science, University of Maryland, Baltimore County, Baltimore, MD.

Kolari, P.; Finin, T.; and Joshi, A. 2006. SVMs for the Blogosphere: Blog Identification and Splog Detection. In Computational Approaches to Analyzing Web Logs: Papers from the 2006 AAAI Spring Symposium. Technical Report SS-06-03. Menlo Park, CA: AAAI Press.

Kolari, P.; Java, A.; and Finin, T. 2006. Characterizing the Splogosphere. Paper presented at the 3rd Annual Workshop on the Web Loggging Ecosystem: Aggregation, Analysis and Dynamics, Edinburgh, Scotland, 23 May.

Kolari, P.; Java, A.; Finin, T.; Oates, T.; and Joshi, A. 2006. Detecting Spam Blogs: A Machine Learning Approach. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading Behavior in Large Blog Graphs. In *Proceedings of the SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Martineau, J.; Java, A.; Kolari, P.; Finin, T.; Joshi, A.; and

AAAI Executive Council Nominations

Every two years, the AAAI membership elects an individual to serve a two-year term as president-elect, followed by two years as president, and, finally, two years as immediate past president. In addition, every year four new councilors are elected to serve three-year terms on the AAAI Executive Council. All elected councilors are expected to attend at least two council meetings per year, and actively participate in AAAI activities. Nominees must be current members of AAAI.

The Nominating Committee encourages all regular members in good standing to place an individual's name before them for consideration. (Student and library members are not eligible to submit candidates' names.) The Nominating Committee, in turn, will nominate one candidate for president-elect and eight candidates for councilor in the spring. In addition to members' recommendations, the committee will actively recruit individuals in order to provide a balanced slate of candidates. AAAI members will vote in the late spring.

To submit a candidate's name for consideration, please send the individual's name, address, phone number, and e-mail address to Carol Hamilton, executive director, AAAI, 445 Burgess Drive, Menlo Park, CA 94025; by fax to 650/321-4457; or by e-mail to hamilton@aaai.org. Nominators should contact candidates prior to submitting their names to verify that they are willing to serve, should they be elected.

The deadline for nominations is November 1, 2008.

Mayfield, J. 2007. Blogvox: Learning Sentiment Classifiers. (Student Abstract). In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence,* 1888–1889. Menlo Park, CA: AAAI Press.

Mishne, G. 2007. Applied Text Analytics for Blogs. Ph.D. Dissertation, Intelligent Systems Lab, University of Amsterdam, Amsterdam, The Netherlands.

Pennock, D. M.; Flake, G. W.; Lawrence, S.; Glover, E. J.; and Giles, C. L. 2002. Winners Don't Take All: Characterizing the Competition for Links on the Web. *Proceedings of the National Academy of Sciences* 99(8): 5207–5211.

Shi, X.; Tseng, B.; and Adamic, L. 2007. Looking at the Blogosphere Topology through Different Lenses. In *Proceedings of the Second International Conference on Web Logs and Social Media*. Menlo Park, CA: AAAI Press.

Tim Finin is a professor of computer science and electrical engineering at the University of Maryland Baltimore County. He has over 35 years of experience in the applications of AI to problems in information systems, intelligent interfaces, and robotics. He holds degrees from MIT and the University of Illinois and has held positions at Unisys, the University of Pennsylvania, and the MIT AI Laboratory.

Anupam Joshi is a University of Maryland Baltimore County professor with research interests in the broad area of networked computing and intelligent systems. He currently serves on the editorial board of the *International Journal of the Semantic Web and Information*.

Pranam Kolari is a senior research engineer at Yahoo! Search Sciences. He received a Ph.D. in computer science. His dissertation was focused on spam blog detection, with tools developed in use both by academe and industry. He has active research interest in internal corporate blogs, the semantic web, and blog analytics.

Akshay Java is a University of Maryland Baltimore County Ph.D. student. His dissertation is on identifying influence and opinions in social media. His research interests include blog analytics, information retrieval, natural language processing, and the semantic web .

Anubhav Kale received an M.S. degree in computer science from the University of Maryland Baltimore County in May 2007. His thesis research demonstrated the effectiveness of detecting sentiment associated with links between blog posts and using this to enhance blog community recognition algorithms. He is currently a software engineer at Microsoft.

Amit Karandikar received an M.S. degree in computer science from the University of Maryland Baltimore County in May 2007. His thesis research produced a generative model for the blogosphere that modeled both the reading and writing activities of bloggers. He is currently a software engineer at Microsoft.