Identifying the Influential Bloggers in a Community

Nitin Agarwal, Huan Liu, Lei Tang
Arizona State University
Tempe, AZ 85287, USA
{Nitin.Agarwal.2, Huan.Liu,
Lei.Tang}@asu.edu

Philip S. Yu
University of Illinois at Chicago
Chicago, IL 60607, USA
psyu@cs.uic.edu

ABSTRACT

Blogging becomes a popular way for a Web user to publish information on the Web. Bloggers write blog posts, share their likes and dislikes, voice their opinions, provide suggestions, report news, and form groups in Blogosphere. Bloggers form their virtual communities of similar interests. Activities happened in Blogosphere affect the external world. One way to understand the development on Blogosphere is to find influential blog sites. There are many non-influential blog sites which form the "the long tail". Regardless of a blog site being influential or not, there are influential bloggers. Inspired by the high impact of the influentials in a physical community, we study a novel problem of identifying influential bloggers at a blog site. Active bloggers are not necessarily influential. Influential bloggers can impact fellow bloggers in various ways. In this paper, we discuss the challenges of identifying influential bloggers, investigate what constitutes influential bloggers, present a preliminary model attempting to quantify an influential blogger, and pave the way for building a robust model that allows for finding various types of the influentials. To illustrate these issues, we conduct experiments with data from a real-world blog site, evaluate multi-facets of the problem of identifying influential bloggers, and discuss unique challenges. We conclude with interesting findings and future work.

Categories and Subject Descriptors: J.4 [Social and Behavioral Science]: Economics, Sociology

General Terms: Algorithm, Design, Experimentation, Human Factors, Measurement, Performance, Verification.

Keywords: Social Networks, Blogosphere, Influential Bloggers.

1. INTRODUCTION

The advent of participatory Web applications (or Web 2.0 [23]) has created online media that turn the former mass information consumers to the present information producers [9]. Examples include blogs, wikis, social annotation and tagging, media sharing, and other such services. A "blog" is a weblog at a website where the entries by individuals are displayed in reverse chronological order. A typical blog can combine text, images, and links to other blogs and to Web pages. These entries can be blog posts or comments - the follow-up posts linked to some specific posts. Blogging is becoming a popular means for mass Web users to express, communicate, share, collaborate, debate, and reflect. Blogosphere is the virtual universe that contains all blogs. Bloggers, the blog writers, loosely form their special interest communities where they share thoughts, express opinions, debate ideas, and offer suggestions interactively. Blogosphere provides a conducive platform to build the *virtual communities* of special interests. It reshapes business models [26], inspires viral marketing [25], provides trend analysis and sales prediction [11, 22], aids counter-terrorism efforts [3] and acts as grassroot information sources [27].

In a physical world, according to [14], 83% of people prefer consulting family, friends or an expert over traditional advertising before trying a new restaurant, 71% of people do the same before buying a prescription drug or visiting a place, and 61% of people talk to family, friends or an expert before watching a movie. In short, before people buy or make decisions, they talk, and they listen to other's experience, opinions, and suggestions. The latter affect the former in their decision making, and are aptly termed as the *influentials* [14]. Influence has always been unabated interest in business and society. As the pervasive presence and ease of use of the Web, an increasing number of people with different backgrounds flock to the Web - a virtual world to conduct many previously inconceivable activities from shopping, to making friends, and to publishing. As we draw parallels between physical and virtual communities, among citizens of the blogosphere, we are intrigued by the questions like whether there exist the influentials in a virtual community (a blog), who they are, and how to find them.

Since the bloggers can be connected in a virtual community anywhere anytime, the identification of the influential bloggers can benefit all in developing innovative business opportunities, forging political agendas, discussing social and societal issues, and lead to many interesting applications. For example, the influentials are often *market-movers*. Since they can influence buying decisions of the fellow bloggers, identifying them can help companies better understand the key concerns and new trends about products interesting to them, and smartly affect them with additional information and consultation to turn them into unofficial spokesmen.

WSDM'08, February 11–12, 2008, Palo Alto, California, USA.

As reported in [5], approximately 64% advertising companies have acknowledged this phenomenon and are shifting their focus toward blog advertising.

The influentials could also *sway* opinions in political campaigns, elections, and affect reactions to government policies [4]. Tapping on the influentials can help understand the changing interests, foresee potential pitfalls and likely gains, and adapt plans timely and pro-actively (not just reactively). The influentials can also help in customer support and troubleshooting since their solutions are trustworthy because of the sense of authority these influentials possess. For example, Macromedia[1] aggregates, categorizes and searches the blog posts of 500 people who write about Macromedia's technology. Instead of going through every blog post, an excellent entry point is to start with the influentials' posts.

Some recent numbers from Technorati[2] show a 100% increase in the size of Blogosphere every six months, "..., about 1.6 Million postings per day, or about 18.6 posts per second"[3]. Blogosphere has grown over 60 times during the past three years. Since new blog posts are being generated with such a blazing fast rate, novel ways have to be developed in order to keep track of everything happening in Blogosphere.

Researchers have studied the influence in Blogosphere about influential *blog sites* [8, 18] (more in Section 2). Regardless of a blog site being influential or not, a multi-authored blog site can have its influential bloggers. Influential bloggers of a site have impact on the fellow bloggers as in a physical community. In this paper, we address a novel problem of identifying *influential bloggers* on a blog site and investigate its issues and challenges.

- Are there influential bloggers as in a physical community? Are they simply active bloggers?
- What measures should be used to define influential bloggers? A solution can be subjective, depending on the need for identifying influential bloggers.
- How to find influential bloggers? As there is no training data to tell us who are influential bloggers or not, it is infeasible to apply classification. Combining the statistics collected for each blogger, can we create a robust model that quantitatively tells how influential a blogger is?
- Can we tune/adjust the model to identify different classes of influential bloggers to satisfy various needs?

In the following, we first review the literature and differentiate this work from the existing ones. In Section 3, we study the statistics collectable from a blog site, and define the problem of identifying influential bloggers. In Section 4, we propose a preliminary model that allows for evaluating different key measures for identifying the influentials and can be adapted to look for different types of influential bloggers. In Section 5, we conduct an empirical study to evaluate many aspects of the proposed approach and its effectiveness, and observe how the key measures work with a correlation study. Finally we conclude our work with some possible future directions in Section 6.

[1]http://weblogs.macromedia.com/
[2]http://technorati.com/
[3]http://www.sifry.com/alerts/archives/000436.html

2. RELATED WORK

Blogosphere expands speedily since its inception. This has attracted a surge of research on Blogosphere. Most studies are conducted in terms of social networks. Emergence of communities in the network can be found frequently at a microscopic level [19]. Researchers work on community detection to explore various communities in Blogosphere [7, 28, 20, 3]. Authors in [2] consider influence a characteristic of virtual communities, among others like membership, reinforcement of needs, shared emotional connection, whose presence governs the establishment of a community. Link structures and overlapping between different sub-communities are used to help identify influence between them.

2.1 Ranking Blogs vs. Webpage Ranking

The problem of ranking blog sites or bloggers differs from that of finding authoritative webpages. As pointed out in [18], *blog sites in the blogosphere are very sparsely linked and it is not suitable to rank blog sites using Web ranking algorithms* like PageRank [24] and HITS [16]. The Random Surfer model of webpage ranking algorithms [24] does not work well for sparsely linked structures. The temporal aspect is most significant in blog domain. While a webpage may acquire authority over time (its adjacency matrix gets denser), a blog post or a blogger's influence diminishes over time. This is due to the fact that the adjacency matrix of blogs (considered as a graph) will get sparser as thousands of new sparsely-linked blog posts appear every day.

Some recent work [18] suggests to add implicit links to increase the density of link information based on topics. If two blogs are talking about the same topic, an edge can be added between these two blogs based on the topic similarity or *information epidemics*. However, constructing links based on the topics still remains an area of research.

2.2 Influential Blog Sites

Finding *influential blog sites* in the blogosphere is an important research problem, which studies how some blog sites influence the external world and within the blogosphere [8]. It is perpendicular to the problem of identifying influential bloggers. Given the nature of the blogosphere, influential blog sites are few. A large number of non-influential sites belong to the long tail [1] where abundant new business, marketing, and development opportunities can be explored. Our work is about identifying influential bloggers at a blog site regardless of the site being influential or not. We briefly review some work on influential blog sites.

Gruhl et al [10] study information diffusion of various topics in the blogosphere from individual to individual, drawing on the theory of infectious diseases. A general cascade model is adopted. They associate 'read' probability and 'copy' probability with each edge of the blogger graph indicating the tendency to read one's blog post and copy it, respectively. They also parameterize the stickiness of a topic which is analogous to the *virulence* of a disease.

An interesting problem related to viral marketing [25, 15] is how to maximize the total influence in the network (of blog sites) by selecting a fixed number of nodes in the network. A greedy approach can be adopted to select the most influential node in each iteration after removing the selected nodes. This greedy approach outperforms PageRank, HITS and ranking by number of citations, and is robust in filtering splogs (spam blogs) [12].

The work discussed in this paper is about *identifying influential bloggers at one blog site* and differs from those briefly reviewed above. A blog site is a special type of social network that contains information such as outlinks (other blog posts it is referring to), inlinks (other blog posts that are citing this blog post), comments which is not present in a general social network. Identifying the influential bloggers at a blog site requires the integrated use of the information specific to a blog site.

Influential bloggers are not necessarily active bloggers at a blog site. Many blog websites list top bloggers or top blog posts in some time frame (e.g., monthly). Those top lists are usually based on some traffic information (e.g., how many posts a blogger posted, or how many comments a blog post received) [8]. Certainly these statistics would leave out those blog sites or bloggers who were not active.

With the speedy growth of the blogosphere, it is increasingly difficult, if at all possible, to manually track the development and happenings in the blogosphere. This work is an effort to help understand the blogosphere. In the following, we first study the concept of influential bloggers at a blog site and then propose a preliminary model of computationally identifying influential bloggers.

3. INFLUENTIAL BLOGGERS

Blogs can be categorized into two major types: *individual* and *community blogs*. Individual blogs are single-authored who record their thoughts, express their opinions, and offer suggestions or ideas. Others can comment on a blog post, but cannot start a new line of blog posts. These are more like diary entries or personal experiences. Examples of individual blogs are Sifry's Alerts: David Sifry's musings[4] (Founder & CEO, Technorati), Ratcliffe Blog–Mitch's Open Notebook[5], The Webquarters[6], etc. A community blog is where each blogger can not only comment on some blog posts, but also start some topic lines. Examples of community blog sites are Google's Official Blog[7], The Unofficial Apple Weblog[8], Boing Boing: A Directory of Wonderful Things[9] etc. For an individual blog, the host is the only one who initiates and leads the discussions and thus is naturally the influential blogger of his/her site. For a community blog where many have equal opportunities to participate, we study who are the influentials in a virtual community. Henceforth, blogs refer to community blogs.

Each blog post is often associated with some metadata like post's author, post annotations, post's date and time, number of comments. In addition, one can also collect certain statistics from the blog website for example, *outlinks* - posts or articles to which the author has referred; *inlinks* - other posts that refer to this post, post length; *average length of comments* per post; and the rate at which comments are posted on a blog post. Since a long blog post can simply contain many outlinks, outlinks are normalized by the length of the blog post. Inlinks are collected using Technorati API[10].

[4]http://www.sifry.com/alerts/
[5]http://www.ratcliffeblog.com/
[6]http://webquarters.blogspot.com/
[7]http://googleblog.blogspot.com/
[8]http://www.tuaw.com/
[9]http://boingboing.net/
[10]http://technorati.com/developers/api/cosmos.html

In the simplest case, one can approximate an influential blogger with an active blogger who posts frequently. Since this is not the case in a physical world where a voluble person is not necessarily or seldom influential, we are inquisitive whether we can employ the above metadata and statistics to identify influential bloggers. The search for influential bloggers boils down to the question on how to define an influential blogger. First, active bloggers are not necessarily influential and influential bloggers can be inactive. Hence, we categorically divide bloggers into four types: active and influential, active and non-influential, inactive and influential, and inactive and non-influential. Second, while active bloggers can be simply defined by how frequently a blogger posts, it is a more complex matter how to define an influential blogger with the aid of the above mentioned statistics.

Recognizing the subjective nature of defining an influence blogger, we propose a preliminary model to quantify the properties of the influential bloggers by combining various statistics collectable from a blog site and assigning influence scores to each blogger and their blog posts. Next, we investigate how these statistics can be used in various ways to adjust the model for different purposes. In this work, we first develop an intuitive model that goes beyond the post frequency and allows the use of the combination of statistics. Then we demonstrate how to use this model to identify influential bloggers who may or may not be active, and further investigate how to further refine and evolve the preliminary model in finding various types of influential bloggers.

An intuitive way of defining an influential blogger is to check if the blogger has any influential blog post, i.e., *A blogger can be influential if s/he has more than one influential blog post.* Assume we have an influence score[11] for a post p_i, $I(p_i)$. For a blogger b_k who has N blog posts, $\{p_1, p_2, ..., p_N\}$, their influence scores can be ranked in descending order, and her influence index, *iIndex(b_k)* can be defined as $\max(I(p_i))$, where $1 \leq i \leq N$. Given a set U of M bloggers, $\{b_1, b_2, ..., b_M\}$, the problem of identifying influential bloggers is defined as determining an ordered subset V of K[12] bloggers, $\{b_{j_1}, b_{j_2}, ..., b_{j_K}\}$ that are ordered according to their *iIndex* such that $V \subseteq U$ and $K \leq M$, i.e. $iIndex(b_{j_1}) \geq iIndex(b_{j_2}) \geq ... \geq iIndex(b_{j_K})$. V containss K most **influential bloggers**. For all the blog posts $\{p_1, p_2, ..., p_L\}$ by all M bloggers, **influential blog posts** are those whose influence scores are greater than $iIndex(b_{j_K})$ or, $I(p_l) \geq iIndex(b_{j_K})$ for $1 \leq l \leq L$. Hence, we have the following corollary: those bloggers who published blog posts that satisfy $I(p_l) \geq iIndex(b_{j_K})$, for $1 \leq l \leq L$ will be called influential bloggers because their *iIndex* will be greater than or equal to $iIndex(b_{j_K})$.

Having formulated the problem of identifying influential bloggers, we now study the intuitive characteristics that help define *iIndex* and I so as to build an experimental model that can gauge the influence to distinguish between "influential" and "activeness" properties of bloggers.

4. IDENTIFYING THE INFLUENTIALS

We first present some desirable properties related to blogpost influence which can be approximately defined by collectable statistics, next propose a preliminary model of identifying the influentials using these statistics, then discuss

[11]These concepts are defined mathematically in Section 4.2.
[12]Note that K is a user specified parameter.

some interesting issues that can be evaluated by experimenting the preliminary model.

4.1 An initial set of intuitive properties

Following [14], one is influential if s/he is recognized by fellow citizens, can generate follow-up activities, has novel perspectives or ideas, and is often eloquent. Below we examine how this initial set of intuitive properties can be approximated by some collectable statistics.

- Recognition - An influential blog post is recognized by many. This can be equated to the case that an influential post p is referenced in many other posts, or its number of inlinks (ι) is large. The influence of those posts that refer to p can have different impact: the more influential the referring posts are, the more influential the referred post becomes.

- Activity Generation - A blog post's capability of generating activity can be indirectly measured by how many comments it receives, the amount of discussion it initiates. In other words, few or no comment suggests little interest of fellow bloggers, thus non-influential. Hence, a large number of comments (γ) indicates that the post *affects* many such that they care to write comments, and therefore, the post can be influential. There are increasing concerns over spam comments that do not add any value to the blog posts or blogger's influence. Fighting spam is outside the scope of this work and recent research can be found in [17, 21].

- Novelty - Novel ideas exert more influence as suggested in [14]. Hence, the number of outlinks is an indicator of a post's novelty. A large number of outlinks (θ) may suggest that a post refers to many other blog posts or articles, indicating that it is less likely to be novel. The number of outlinks is negatively correlated with the number of comments which means more outlinks reduces people's attention. This is confirmed later in Section 5.2.5.

- Eloquence - An influential is often eloquent [14]. This property is most difficult to approximate using some statistics. Given the informal nature of the blogosphere, there is no incentive for a blogger to write a lengthy piece that bores the readers. Hence, a long post often suggests some necessity of doing so. Therefore, we use the length of a post (λ) as a heuristic measure for checking if a post is influential or not. The blog post length is positively correlated with number of comments which means longer blog posts attract people's attention. This is confirmed later in Section 5.2.5.

The above four form an initial set of properties possessed by an influential post. There are certainly some other potential properties. It is also evident that each of the above four may not be sufficient on its own, and they should be used jointly in identifying influential bloggers. For example, a high θ and a poor λ could identify a messenger blog post. Starting with this initial set, we next build a preliminary model that allows us to examine, analyze, modify, and extend the model.

4.2 Influence graph - a preliminary model

Blog-post influence can be visualized in terms of an influence graph or *i-graph* in which the influence of a blog post flows among the nodes. Each node of an i-graph represents a single blog post characterized by the four properties (or parameters): ι, θ, γ and λ. i-graph is a directed graph with ι and θ representing the incoming and outgoing influence flows of a node, respectively. Hence, if I denotes the influence of a node (or blog post p), then $InfluenceFlow$ across that node is given by,

$$InfluenceFlow(p) = w_{in} \sum_{m=1}^{|\iota|} I(p_m) - w_{out} \sum_{n=1}^{|\theta|} I(p_n) \quad (1)$$

where w_{in} and w_{out} are the weights that can be used to adjust the contribution of incoming and outgoing influence, respectively. p_m denotes all the blog posts that link to the blog post p, where $1 \leq m \leq |\iota|$; and p_n denotes all the blog posts that are referred by the blog post p, where $1 \leq n \leq |\theta|$. $|\iota|$ and $|\theta|$ are the total numbers of inlinks and outlinks of post p. $InfluenceFlow$ measures the difference between the total incoming influence of all inlinks and the total outgoing influence by all outlinks of the blog post p. $InfluenceFlow$ accounts for the part of influence of a blog post that depends upon inlinks and outlinks. From Eq. 1, it is clear that the more inlinks a blog post acquires the more recognized it is, hence the more influential it gets; and an excessive number of outlinks jeopardizes the novelty of a blog post which affects its influence.

As discussed earlier, the influence (I) of a blog post is also proportional to the number of comments (γ_p) posted on that blog post. We can define the influence of a blog post, p as,

$$I(p) \propto w_{com}\gamma_p + InfluenceFlow(p) \quad (2)$$

where w_{com} denotes the weight that can be used to regulate the contribution of the number of comments (γ_p) towards the influence of the blog post p. We consider an additive model because additive function is good to determine the combined value of each alternative [6]. It also supports preferential independence of all the parameters involved in the final decision. Since most decision problems like the one at hand are multi-objective, a way to evaluate trade-offs between the objectives is needed. A weighted additive function can be used for this purpose [13].

From the discussion in Section 4.1, we consider blog post quality as one of the parameters that may affect influence of the blog post. Although there are many measures that quantify the goodness of a blog post such as fluency, rhetoric skills, vocabulary usage, and blog content analysis[13], for the sake of simplicity, we here use the length of the blog post as a heuristic measure of the goodness of a blog post in the context of blogging. We define a weight function, w, which rewards or penalizes the influence score of a blog post depending on the length (λ) of the post. The weight function could be replaced with appropriate content and literary analysis tools. Combining Eq. 1 and Eq. 2, the influence of a blog post, p, can thus be defined as,

$$I(p) = w(\lambda) \times (w_{com}\gamma_p + InfluenceFlow(p)) \quad (3)$$

[13]A reason we did not adopt any of these is their computation is beyond the scope of this work. We use some simpler measure to examine its effect in determining influence.

The above equation gives an influence score to each blog post. Note that the four weights can take more complex forms and can be tuned. We will evaluate and discuss their effects further in the empirical study.

Now we consider how to use I to determine whether a blogger is influential or not. According to the definition of influential blogger in Section 3, a blogger can be considered influential if s/he has at least one influential blog post. We use the blog post with maximum influence score as the representative[14] and assign its influence score as the *blogger influence index* or *iIndex*. For a blogger B, we can calculate the influence score for each of B's N posts and use the maximum influence score as the blogger's *iIndex*, or

$$iIndex(B) = \max(I(p_i)) \quad (4)$$

where $1 \leq i \leq N$. With *iIndex*, we can rank bloggers on a blog site. The top k among the total bloggers are the most influential ones. Thresholding is another way to find influential bloggers whose *iIndices* are greater than a threshold. However, determining a proper threshold is crucial to the success of such a strategy and requires more research.

4.3 Issues of identifying the influentials

The preliminary model presents a palpable way of identifying influential bloggers and allows us to address many relevant issues such as evaluation, feasibility, efficacy, subjectivity, and extension.

- Can we use this model to differentiate influential bloggers from active bloggers? We study the existence of influential bloggers at a blog site by applying the preliminary model.
- How can we evaluate the model's performance in identifying the influential bloggers? Are influential blog posts indeed different from non-influential blog posts?
- How can we properly determine the weights when combining the four parameters in *iIndex*? If one changes the value of a weight, will the change significantly affect the ranking of influential bloggers? How these weights can help find special influential bloggers?
- How do we handle the subjectivity aspect of the problem of identifying influential bloggers as different people may have disparate preferences? Since we have access to the whole history of the blog site, we look into these questions by consecutively studying the influentials in multiple 30-day windows. Can we also employ the model to find any temporal patterns of the influential bloggers?
- Are all the four parameters necessary? We design and perform a correlation study. Some of the parameters may be correlated with each other, so one of them may be redundant. Pairwise correlation analysis is thus conducted.
- How can we extend the preliminary model? Are there any other parameters that can be incorporated in a refined model?

[14]There could be other ways. For example, if one wants to differentiate a productive influential blogger from non-prolific one, one might use another measure.

In the next, we set out to use the proposed model in an empirical study, attempt to experimentally address these issues, report preliminary results, and suggest new lines of research in finding influential bloggers.

5. FURTHER STUDY & EXPERIMENTS

We first discuss the need for experimental data, and select a real-world blog site for experiments; and second, we design various experiments with the preliminary model using *iIndex*, and answer the questions raised in Section 4.3 based on the experimental results. In the process, we develop and elaborate an evaluation procedure for effective comparison.

5.1 Data collection

Data collection is one of the critical tasks in this work. To our best knowledge, our effort is the first attempt to find influential bloggers. Hence, there are no available blog data sets for the purposes of our experiments. We need to collect real-world data.

There exist many blog sites. Some like Google's Official Blog site act as a notice board for important announcements rather than for discussions, sharing opinions, ideas and thoughts; some do not provide most of the statistics needed in our work, although they can be obtained via some additional work (more explanation later). A few publicly available blog datasets like the BuzzMetric dataset[15] were designed for different research experiments so there is no way to obtain some key statistics required in this work.

Therefore, we crawled a real-world blog site that provides the most statistics required in our experiments. The advantages of of doing so include (1) minimizing our effort on figuring out ways to obtain the needed statistics, and (2) maximizing the reproducibility of our experiments independently. The Unofficial Apple Weblog (TUAW) site is such a site that satisfies these requirements. This blog site provides most needed information like blogger identification, date and time of posting, number of comments, and outlinks. The only missing piece of information at TUAW is the *inlinks* information, which we can obtain using Technorati API[16]. We crawled the TUAW blog site and retrieved all the blog posts published since it was set up. We have collected over $10,000$ posts till January 31, 2007. We keep the complete history of the TUAW blog site and update it incrementally. All the statistics obtained after crawling is stored in a relational database for fast retrieval later[17].

5.2 Results and discussions

The following subsections introduce the experiments, results, and discussions corresponding to the questions raised in the Section 4.3.

5.2.1 Influential Bloggers and Active Bloggers

Many blog sites publish a list of top bloggers based on their activities on the blog site. The ranking is often made according to the number of blog posts each blogger submitted over a period of time. In this paper, we call these people *active* bloggers. Since the top bloggers on the blog site TUAW are those from the last 30 days, we define our

[15]http://www.nielsenbuzzmetrics.com/cgm.asp
[16]http://technorati.com/developers/api/cosmos.html
[17]This dataset will be made available upon request for research purposes.

Top 5 TUAW Bloggers	Top 5 Influential Bloggers
Erica Sadun	*Erica Sadun*
Scott McNulty	Dan Lurie
Mat Lu	*David Chartier*
David Chartier	*Scott McNulty*
Michael Rose	Laurie A. Duncan

Table 1: Two lists of the top 5 bloggers according to TUAW and our model, respectively.

study window of 30 days as well. Using the number of posts of a blogger posted is obviously an oversimplified indicator, which basically says the most frequent blogger is an influential one. Such a status can be achieved by simply submitting many posts, as even junk posts are counted. Hence, an active blogger may not be an influential one; and in the same spirit, an influential blogger need not be an active one. In other words, the most active k bloggers are not necessarily the top influential one, and an inactive blogger can still be an influential one.

In our first experiment, we generate a list of top-k bloggers using the preliminary model proposed in Section 3. We set the default values of all the weights as 1 assuming they are equally important. An in-depth study of these weights is in Section 5.2.2. By setting $k = 5$, we compare the top 5 influential bloggers with the top 5 bloggers published at TUAW. Table 1 presents two lists of top 5 bloggers according to TUAW and based on the proposed model using *iIndex*: the first column contains the top 5 bloggers published by TUAW and the second column lists the top 5 influential bloggers. Names in *italics* are the bloggers present in both lists. Three out of 5 TUAW top bloggers are also among the top 5 influential bloggers identified by our model. This set of bloggers suggests that some of the bloggers can be both active and influential. Some active bloggers are not influential and some influential bloggers are not active. For instance, 'Mat Lu' and 'Michael Rose' in the TUAW list, so they are active; and 'Dan Lurie' and 'Laurie A. Duncan' in the list of the influentials, but they are not active.

In total, there could be four types of bloggers: both active and influential, active but non-influential, influential but inactive, inactive and non-influential. Since we have all the needed statistics, we can delve into the numbers and scrutinize their differences of the first three groups of bloggers. Their detailed statistics are presented in Table 2. *Inactive and non-influential bloggers* seldom submit blog posts and submitted posts do not influence others, so this group does not show up in Table 2.

- *Active and influential bloggers* who actively post and some of them are influential posts. 'Erica Sadun', 'David Chartier' and 'Scott McNulty' are of this category. This can be verified by the large number of posts and the large number of comments and citations by other bloggers. For instance, 'Erica Sadun' submitted 152 posts in the last 30 days, among which 9 of them are influential, attracting a large number of readers evidenced by 75 comments and 80 citations.

- *Inactive but influential bloggers.* These bloggers submit a few but influential posts. 'Dan Lurie' published only 16 posts (much fewer than 152 posts comparing with 'Erica Sadun', an active influential blogger) in the last 30 days. Dan was not selected by TUAW as a top blogger. A closer look at his blog posts reveals that 4 of his blog posts are influential, i.e., 25% of the blog posts by 'Dan Lurie' are influential. One of his influential posts is about iPhone[18], which attracted a large number of bloggers to comment and intrigued a heated discussion of the new product (77 comments and 33 inlinks). Its length is 1417 bytes, and there are no outlinks. All these numbers suggest that the post is detailed, innovative, and interesting to other bloggers. By reading the content, we notice that the post is a detailed account of his personal experience rather than extracts from external news sources. This kind of posts allows a reader to experience something new, thus often results in many comments and discussions.

- *Active but non-influential bloggers.* These bloggers post actively, but their posts may not generate sufficient interests to be ranked as the top 5 influentials. 'Mat Lu' and 'Michael Rose' were ranked 3^{rd} and 4^{th} top bloggers by TUAW, as they submitted 73 and 58 blog posts in the last 30 days (around 2 posts a day), respectively. Though these are much more than the 16 posts of 'Dan Lurie', they are not among the top 5 influential bloggers because their other statistics are not comparable with those of the influentials (i.e., having fewer comments and inlinks, and more outlinks).

A closer look at two influential blog posts.

Here we further study the most influential blog posts by number one ('Erica Sadun') and number five ('Laurie A. Duncan') influential bloggers, respectively. The most influential blog post by 'Erica Sadun' is on keynote speech of Apple Inc. CEO, Steve Jobs[19] which fostered overwhelming discussions through 63 comments and 80 inlinks. By reviewing the comments, we observe that most people appreciated her efforts and found the blog post extremely informative. The blog post was the first one dispensing a minute-by-minute description of the much-awaited keynote speech, new products, and services Apple would launch. The blog post was well-written and did not borrow information from any other sources. The most influential blog post by 'Laurie A. Duncan' detailed the violation of license agreements by macZOT[20] with a developer[21]. This incident instigated a lot of discussion through 57 comments and 20 inlinks. Many people commented and cited this blog post, and agreed with the miserable state of license agreements, being appalled by how big companies could exploit small developers by finding loopholes in the laws. Similar sentiments expressed in a surge of comments are an important feature of many influential blog posts. The above study of two most influential posts shows the efficacy of the proposed model.

5.2.2 Evaluating the Model

As we know, there is no training and testing data for us to evaluate the efficacy of the proposed model. The absence of ground truth about influential bloggers presents another

[18]http://www.tuaw.com/2007/01/09/iphone-will-not-allow-user-installable-applic ations/

[19]http://www.tuaw.com/2007/01/09/macworld-2007-keynote-liveblog/

[20]http://www.maczot.com/

[21]http://www.tuaw.com/2007/01/04/xpad-developer-says-maczot-and-brian-ball-ripped-him-off/

		Number of Comments		Number of Inlinks		Length of Blog Post		Number of Outlinks		Total Number	Influential
		Max	Average	Max	Average	Max	Average	Max	Average	of Blog Posts	Blog Posts
Active + Influential	Erica Sadun	75	11.0197	80	10.1316	2935	830.0066	15	2.5329	152	9
	David Chartier	56	11.3088	32	10.25	3529	1054.912	14	4.3529	68	4
	Scott McNulty	112	11.5607	33	8.9252	2246	623.2991	12	2.5888	107	3
Inactive + Influential	Dan Lurie	96	19.6316	37	10.2632	1569	793.7368	4	2.3158	16	4
	Laurie A. Duncan	65	16.2895	34	10.6053	2888	993.8947	11	3.4737	26	2
Active + Non-Influential	Mat Lu	42	8.0294	29	10.0147	1699	771.1471	12	4.1029	73	0
	Michael Rose	31	8.7273	21	9.6061	1378	735.9848	15	6.1515	58	0

Table 2: Comparison of statistics between different bloggers.

Bloggers	Active	Inactive
Influential	S1: 17	S2: 7
Non-influential	S3: 3	S4: 0/1

Table 3: Intersection of Digg and top 20 from our model.

Bloggers	Active	Inactive
Influential	S1: 71	S2: 14
Non-influential	S3: 8	S4: 7

Table 4: Distribution of 100 Digg blog posts.

Bloggers	Active	Inactive
Influential	S1: 327	S2: 42
Non-influential	S3: 131	S4: 35

Table 5: Distribution of 535 TUAW blog posts.

challenge. The key issue is how to find a reasonable reference point for which four different types of bloggers can be evaluated so that we can observe their tangible differences. As an alternative to the ground truth, we resort to another Web2.0 site Digg (http://www.digg.com/) to provide a reference point. According to Digg, "Digg is all about user powered content. Everything is submitted and voted on by the Digg community. Share, discover, bookmark, and promote stuff that's important to you!". As people read articles or blog posts, they can give their votes in the form of digg and these votes are recorded on Digg servers. This means, blog posts that appear on Digg are liked by their readers. The higher the digg score for a blog post is, the more it is liked. In a way, Digg can be considered as a large online user survey. Though only submitted blog posts are voted, Digg offers a way for us to evaluate the blog posts of the four types. Digg provides an API to extract data from their database for a window of 30 days. We used this API to obtain the data for the month of January 2007. Given the nature of Digg, a not-liked blog post will not be submitted thus will not appear in Digg. For January 2007, there were in total 535 blog posts submitted on TUAW. As Digg only returns top 100 voted posts, we use these 100 blog posts at Digg as our benchmark in evaluation.

We take the four categories of bloggers, viz. 1. Active and Influential, 2. Inactive and Influential, 3. Active and Non-influential, and 4. Inactive and Non-influential and categorize their posts into S1, S2, S3, and S4, respectively. We rank the blog posts of each category based on the influence score and pick top 20 blog posts from each of the first three categories. We randomly pick 20 blog posts from the last category in which bloggers are neither active nor influential. Next we compare these four sets of 20 blog posts with the Digg set of 100 blog posts to see how many posts in each set also appear in the Digg set. The results are shown in Table 3. From the table, we can see that S1 has 17 out of 20 in the Digg set, and S4 has 0 or 1 found in the Digg set depending on randomization. The results show the differences among the four categories of bloggers and our model identifies the influentials whose blog posts are more liked than others according to Digg. For reference purposes, we also provide the distributions of 100 Digg and 535 TUAW blog posts in Tables 4 and 5, respectively. Note that we selected top 5 active and 5 influential bloggers (Table 1), in which 3 are both active and influential (Table 2). We observe from Tables 3, 4 and 5 that influential bloggers are more likely to be liked than active bloggers. More detailed discussion is omitted due to space limit. An interesting case is about S4 in Table 4 which has 7 blog posts liked by people even though they were non-influential and inactive. A closer examination reveals that one of the bloggers in S4 was ranked 6th in the list of influential bloggers and 4 of his blog posts appeared in Digg. In other words, increasing the number of top influential bloggers will change the current distribution.

5.2.3 *Influential vs. Non-Influential Blog Posts*

Here we study the contrast in the characteristics between influential and non-influential blog posts. Using the definition of influential blog posts from Section 3, we pick influential blog posts submitted by the influential bloggers listed in Table 1. Rest of the blog posts are treated as non-influential blog posts. Totally we have 22 influential and 513 non-influential blog posts for January 2007. Similar to Table 2, we compare the max and average statistics for all the four parameters (comments, inlinks, blog post length, and outlinks) for both influential and non-influential blog posts and report the results in Table 6. It shows influential blog posts are much longer in length and have far more comments. There are a lot more inlinks in influential blog posts, but the number of outlinks is a weaker piece of evidence, though the influential blog posts have slightly smaller number of outlinks.

5.2.4 *Effects and usages of weights*

There are four weights in our preliminary model to regulate the contribution of four parameters toward the calculation of the influence score using Eq 1 & Eq 3. To recall, w_{in} is for the influence from incoming links, w_{out} for the influence from outgoing links, $w(\lambda)$ for the "goodness" of a blog post, and w_{comm} for the number of comments. All weights take real values in $[0, 1]$. We now study how the change of their values will affect the ranking of the influentials.

One may notice that $w(\lambda)$ simply scales the influence score of a blog post, so varying $w(\lambda)$ is not expected to affect the ranking of influential bloggers, but to scale up or down the influence scores. This is verified by conducting experiments in which the other three weights are fixed and only $w(\lambda)$ is varied. We observe that the relative ordering of the influential bloggers remain the same while their influence score is scaled up or down. Although this weight is immaterial for identifying the influentials at one blog site, it can be used in comparing the influential bloggers of different blog sites for

	Number of Comments		Number of Inlinks		Length of Blog Post		Number of Outlinks		Total Number of Blog Posts
	Max	Avg	Max	Avg	Max	Avg	Max	Avg	
Influential Blog Posts	112	74.18	80	38.63	3529	1999.32	15	3.36	22
Non-influential Blog Posts	69	10.84	39	8.96	1930	703.74	27	4.3	513

Table 6: Comparison of statistics between Influential and non-influential blog posts.

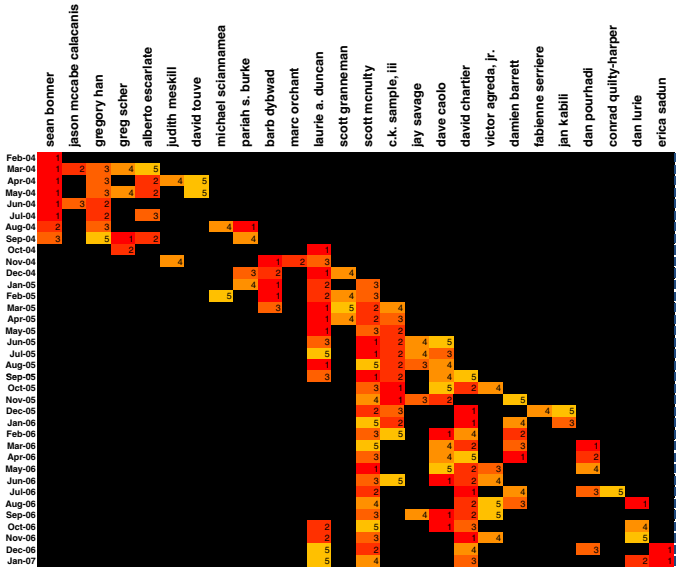


Figure 1: Influential Bloggers' blogging behavior over the whole TUAW blog history.

normalization purposes (outside the scope of this work).

For the remaining three weights, w_{comm}, w_{in} and w_{out}, we fix two of them and observe how the ranking changes by varying the third weight. Fixing w_{in} and w_{out} and varying w_{comm} from 0.0 to 1.0 in steps of 0.1, we observe that the model stabilizes for $w_{comm} \geq 0.6$, i.e., it does not change the ranking of the influential bloggers. While varying w_{in} and w_{out} respectively, we observe that the model stabilizes when $w_{in} \geq 0.9$ and $w_{out} \geq 0.2$. To summarize, we obtain the same ranking of influential bloggers as shown in the right column of Table 1 for $w_{comm} \geq 0.6, w_{in} \geq 0.9, w_{out} \geq 0.2$.

Clearly, the value change of the above three weights can lead to different rankings. This allows one to adjust the weights of the model to attain different goals. With the preliminary model of default setting, we can tune these weights in identifying influential bloggers with different characteristics. For example, by setting w_{in} and w_{out} to 0, we can obtain influential bloggers based on the number of comments a blogger's post obtained. Similarly we can obtain the blog post that received most citations or the blog post including the least outlinks. If one wants to emphasize one aspect, one can tune weights and obtain ranking to reflect that aspect. The increase of w_{out} is one way to discourage the citations of other blog posts, in a way, encouraging a post with independent ideas. In short, these weights provide a means to further evolve and expand the preliminary model for a wide range of applications.

5.2.5 *Temporal patterns of the influentials*

Above, we study the influential bloggers with a time window of 30 days (or monthly). For a blog site that has a reasonably long history, we can also study the temporal patterns of its influential bloggers. The blog site TUAW provides blogging data since its inception February 2004. We hence apply our model to identify top 5 influential bloggers with a moving 30-day window until January 2007, and there is no overlap between two consecutive windows. In total, there are 26 influential bloggers during Feb.2004-Jan.2007. The temporal patterns of the influentials can be observed from a matrix in Figure 1. Influential bloggers are ordered according to the time they were recognized as influential vertically(column-wise), and the rows represent the progression of time. The (i, j)-th cell in this matrix stores the rank of the j^{th} blogger in the i^{th} time window. For example, the first cell (*sean bonner*, *Feb-04*) shows that *Sean Bonner* was ranked top 1 among the influential bloggers list in February 2004[22]. Black cells represent that the particular blogger was not among the top 5 for that time period. The color gradient represents rank of a influential blogger, a darker color representing a better rank.

We can observe some different temporal patterns for the influentials in Figure 1. Among all the 26 bloggers, 17 are influential for at least 4 months. We broadly categorize the influential bloggers into the following:

Long-term influentials They steadily maintain the status of being influential for a very long time. *Scott McNulty* is the best example of this category: *Scott McNulty* is steadily influential from Jan-05 till Jan-07. They can be considered "authority" in the community.

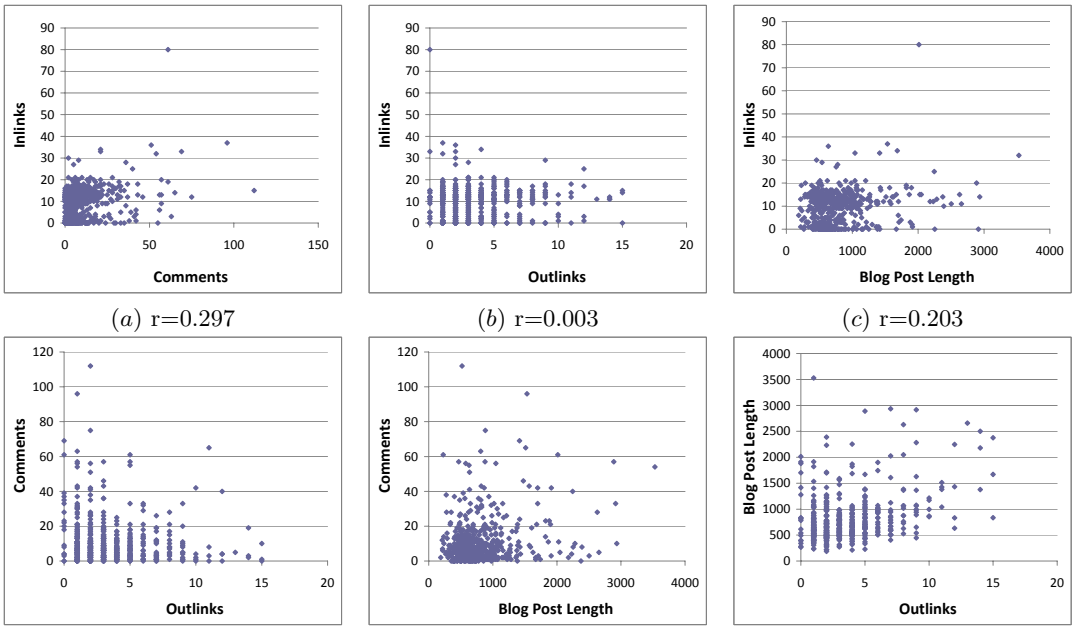
Average-term influentials They maintain their influence status for 4-5 months. Examples of such bloggers from Figure 1 are "Sean Bonner", "Gregory Han", and "Barb Dybward".

Transient influentials They are influential for a **very** short time period (only one or two months). Examples are *Michael Sciannamea*, *Fabienne Serriere* and *Dan Pourhadi*. For instance, *Fabienne Serriere* was influential in Jan-06 and never became influential again.

Burgeoning influentials They are emerging as influential bloggers recently. Bloggers that belong to this category are *Dan Lurie* and *Erica Sadun*. They are the influentials worthy more follow-up examinations.

Disparate bloggers can present different temporal patterns. Long-term influentials are more influential than other bloggers as they are more trustworthy as compared to other bloggers based on a long time of history. Burgeoning influentials have potential to become long-term ones. But it is difficult to say these things about transient influentials as they might become influential by chance. Certainly, there could be many other temporal patterns depending on a particular application. The categories presented here are some examples. Many potential applications can be developed using categories. When we want to know about a new blog site, the best way to approach it is to look at its long-term influentials as they have lasting influence in the community.

[22]In early stage of the blog site, there are a few cases in which there was little blogging activity such as *Feb-04*, *Oct-04*, and *Nov-04*, resulting in fewer than 5 influentials.



(a) r=0.297 (b) r=0.003 (c) r=0.203

(d) r=-0.111 (e) r=0.208 (f) r=0.351

Figure 2: Pairwise correlation plots of the four parameters (ι, θ, λ, and γ) of the blog posts.

The blog posts of those average-term influentials can be used to understand the changing topics. The blog posts of burgeoning influentials might contain the trendy buzz. With accumulated blogging data, we can also learn to predict if a burgeoning influential will more likely become long-term, average-term, or transient.

5.2.6 *Further Experiments*

We conduct more experiments to (1) examine the pairwise correlations of the four factors; and (2) study another statistics - the rate of comments to extend our model.

Correlation analysis. We perform pairwise correlation analysis between the parameters to further examine whether there is any redundant parameter. With four parameters, there are 6 pairwise correlations as shown in Figure 2(a)-(f). The number below each scatter plot is the correlation coefficient. We observe that there is no strong correlation between any pair of parameters. In other words, none of the parameters can be covered by another one. We notice that 5 of 6 scatter plots show positive correlations, but the (d) scatter plot shows some negative correlation, which suggests that more outlinks in a blog post somehow mean fewer comments the post receives, and vice versa. This supports that links among blog posts are different from web links (Section 2).

Rate of comments. This parameter seems a good indicator on how influential a post is. If a post receives many comments in a short period (i.e., it exhibits a spike), it has apparently generated a lot of response, indicating that the post is potentially influential. However, is the opposite true too, i.e., the observation of a flat distribution of comment rates of a blog post implies a non-influential post? We conduct a case study and present the results in Figures 3 and 4 with comment rates of two influential blog posts: one related to the newly publicized iPhone release and the other about a competition held at Apple Inc. Figure 3 exhibits a spiky type of user response. Most of the comments were submitted during the first hour (over 50) after the blog post was published. On the other hand, comment rates in Figure 4 are relatively "flat", around 10 comments per hour even after 7 or 8 hours of the blog post submission. Since the spiky pattern is not a necessary characteristic of an influential post, more research is needed to explore how to incorporate the comment rate. We envision that this parameter can be used to build a more refined model for special time-critical applications like disaster prevention and management, emergency handling.

Other extensions to the preliminary model include 1). study of spam comments filtering to prevent spam attacks using techniques mentioned in [17, 21], 2). study more appropriate blog post quality estimation techniques involving content and literary analysis, and 3). study different functions to non-linearly penalize influence due to outlinks. This basically means assigning negligibly small penalty if few outlinks are present and very high penalty for outrageous number of outlinks. This is required to avoid penalizing those novel blog posts that refer to a few blog posts to support their explanation. One such function could be exponential which would replace $w_{out} \sum_{n=1}^{|\theta|} I(p_n)$ in Eq. 1 with $\exp(w_{out} \sum_{n=1}^{|\theta|} I(p_n))$. We would have to investigate thoroughly the role of w_{out} in such a scenario.

6. CONCLUSIONS AND FUTURE WORK

Blogosphere is one of the fastest growing, social media. The virtual communities in the blogosphere are not constrained by physical proximity and allow for a new form of efficient communications. The influential bloggers naturally exert their influence on other members, lead trends, and affect group interests in a community. They are the conduits of information in their communities. With many great successes of Web 2.0 applications, more and more people take

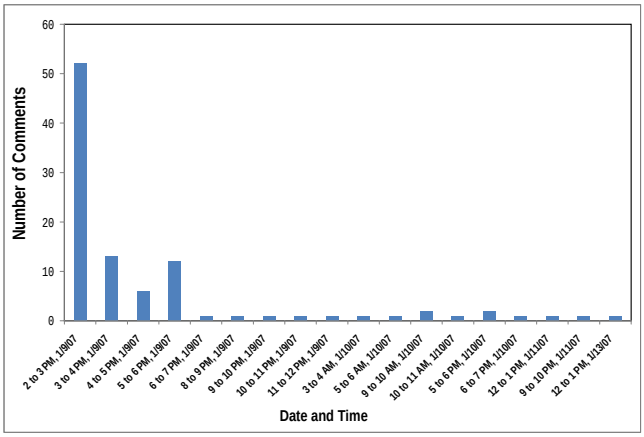


Figure 3: Spiky comments reaction on a blog post related to iPhone.

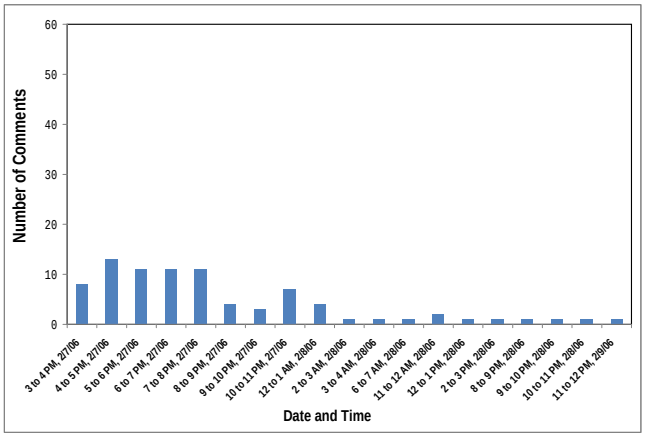


Figure 4: "Flat" comments reaction on a blog post related to some competition in Apple Inc.

part in one form or another of activities in virtual communities. Finding the influential bloggers will not only allow us to better understand interesting activities happening in a virtual world, but also present unique opportunities for industry, sales, and advertisements. With the speedy expansion of the blogosphere, it is vital to develop novel tools that facilitate people to participate, connect, and explore.

We address a novel problem of identifying influential bloggers at a blog site by presenting a preliminary model of identifying influential bloggers of a community blog site. Our work differs from existing works on blogosphere influence over traditional media, influential blog sites, and influence maximization within the blogosphere. Influential bloggers can exist at many blog sites, regardless of these sites being influential or not. We examine essential issues of identifying influential bloggers, evaluate the effects of various collectable statistics from a blog site on determining blog-post influence, develop unique experiments using another Web2.0 application, and conduct experiments by using the whole history of blog posts of a real-world blog site. The extensive but still preliminary work demonstrates that (1) influential bloggers are not necessarily active bloggers, (2) our model can effectively find influential bloggers, (3) by tuning the weights associated with the parameters of the preliminary model, one can examine how different parameters impact the influence ranking for different needs, and (4) the preliminary model can serve as a baseline in identifying influential bloggers and can be extended by incorporating additional parameters to discover different patterns. We expect that the preliminary model will evolve to address many new needs arising from the real (or rather virtual) world.

7. REFERENCES

[1] Chris Anderson. *The long tail : why the future of business is selling less of more*. New York : Hyperion, 2006.

[2] Alvin Chin and Mark Chignell. A social hypertext model for finding community in blogs. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 11–22, New York, NY, USA, 2006. ACM Press.

[3] T. Coffman and S. Marcus. Dynamic classification of groups through social network analysis and HMMs. In *Proceedings of IEEE Aerospace Conference*, 2004.

[4] Daniel Drezner and Henry Farrell. The power and politics of blogs. In *American Political Science Association Annual Conference*, 2004.

[5] T. Elkin. Just an online minute... online forecast. http://publications.mediapost.com/index.cfm?fuseaction =Articles.showArticle art aid=29803.

[6] Gerald D. Fensterer. *Planning and Assessing Stability Operations: A Proposed Value Focus Thinking Approach*. PhD thesis, Air Force Institute of Technology, 2007.

[7] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of web communities. In *6th International Conference on Knowledge Discovery and Data Mining*, 2000.

[8] Kathy E. Gill. How can we measure the influence of the blogosphere? In *Proceedings of the WWW'04: workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[9] Dan Gillmor. *We the Media: Grassroots Journalism by the People, for the People*. O'Reilly, 2006.

[10] D. Gruhl, David Liben-Nowell, R. Guha, and A. Tomkins. Information diffusion through blogspace. *SIGKDD Exploration Newsletter*, 6(2):43–52, 2004.

[11] Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak, and Andrew Tomkins. The predictive power of online chatter. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87, New York, NY, USA, 2005. ACM Press.

[12] Akshay Java, Pranam Kolari, Tim Finin, and Tim Oates. Modeling the spread of influence on the blogosphere. In *Proceedings of the 15th International World Wide Web Conference*, 2006.

[13] R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Cambridge University Press, 1993.

[14] Ed Keller and Jon Berry. *One American in ten tells the other nine how to vote, where to eat and, what to buy. They are The Influentials*. The Free Press, 2003.

[15] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the KDD*, pages 137–146, New York, NY, USA, 2003. ACM Press.

[16] J. Kleinberg. Authoritative sources in a hyperlinked environment. In *9th ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[17] P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[18] Apostolos Kritikopoulos, Martha Sideri, and Iraklis Varlamis. Blogrank: ranking weblogs based on connectivity and similarity features. In *AAA-IDEA '06: Proceedings of the 2nd international workshop on Advanced architectures and algorithms for internet delivery and applications*, page 8, 2006.

[19] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the Bursty Evolution of Blogspace. In *Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[20] Yu-Ru Lin, Hari Sundaram, Yun Chi, Jun Tatemura, and Belle Tseng. Discovery of blog communities based on mutual awareness. In *Proceedings of the 3rd annual workshop on webloging ecosystem: aggreation, analysis and dynamics*, 2006.

[21] Yu-Ru Lin, Hari Sundaram, Yun Chi, Junichi Tatemura, and Belle L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web (AIRWeb)*, pages 1–8, New York, NY, USA, 2007. ACM Press.

[22] Gilad Mishne and Maarten de Rijke. Deriving wishlists from blogs show us your blog, and we'll tell you what books to buy. In *Proceedings of the 15th international conference on World Wide Web*, pages 925–926, New York, NY, USA, 2006. ACM Press.

[23] Tim O'Reilly. What is Web 2.0 - design patterns and business models for the next generation of software. http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, September 2005.

[24] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[25] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 61–70, New York, NY, USA, 2002. ACM Press.

[26] Robert Scoble and Shel Israel. *Naked conversations : how blogs are changing the way businesses talk with customers*. John Wiley, 2006.

[27] Mike Thelwall. Bloggers under the London attacks: Top information sources and topics. In *Proceedings of the 3rd annual workshop on webloging ecosystem: aggreation, analysis and dynamics*, 2006.

[28] Ying Zhou and Joseph Davis. Community discovery and analysis in blogspace. In *Proceedings of the 15th international conference on World Wide Web*, pages 1017–1018, New York, NY, USA, 2006. ACM Press.