Soc. Netw. Anal. Min. (2012) 2:139–162
DOI 10.1007/s13278-011-0039-3

ORIGINAL ARTICLE

Modeling blogger influence in a community

Nitin Agarwal · Huan Liu · Lei Tang · Philip S. Yu

Received: 6 July 2010 / Revised: 18 August 2011 / Accepted: 28 August 2011 / Published online: 6 October 2011

Abstract Blogging has become a popular and convenient way to communicate, publish information, share preferences, voice opinions, provide suggestions, report news, and form virtual communities in the Blogosphere. The blogosphere obeys a power law distribution with very few blogs being extremely influential and a huge number of blogs being largely unknown. Regardless of a (multi-author) blog being influential or not, there are influential bloggers. However, the sheer number of such blogs makes it extremely challenging to study each one of them. One way to analyze these blogs is to find influential bloggers and consider them as the community representatives. Influential bloggers can impact fellow bloggers in various ways. In this paper, we study the problem of identifying influential bloggers. We define influential bloggers, investigate their characteristics, discuss the challenges with identification, develop a model to quantify their influence, and pave the way for further research leading to more sophisticated models that enable categorization of various types of influential bloggers. To highlight these issues, we conduct experiments using data from blogs, evaluate multiple facets of the problem, and present a unique and objective evaluation strategy given the subjectivity in defining the influence, in addition to various other analytical capabilities. We conclude with interesting findings and future work.

Keywords Social network · Blogosphere · Influence · Influential bloggers · Evaluation

N. Agarwal (✉)
University of Arkansas at Little Rock, Little Rock, AR, USA
e-mail: nxagarwal@ualr.edu

H. Liu
Arizona State University, Tempe, AZ, USA
e-mail: Huan.Liu@asu.edu

L. Tang
Yahoo! Labs, Santa Clara, CA, USA
e-mail: ltang@yahoo-inc.com

P. S. Yu
University of Illinois at Chicago, Chicago, IL, USA
e-mail: psyu@cs.uic.edu

1 Introduction

The advent of participatory Web applications (or Web 2.0 (O'Reilly 2005)) has created online media that turn the former mass information consumers to the present information producers (Gillmor 2006). Examples include blogs, wikis, collaborative tagging, media sharing, and other such services. A blog site or simply blog (short for web log) is a collection of entries by individuals displayed in reverse chronological order. These entries, known as the blog posts, can typically combine text, images, and links to other blogs, blog posts, and/or to Web pages. Blogging is becoming a popular means for mass Web users to express, communicate, share, collaborate, debate, and reflect. The 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. The blogosphere provides a platform conducive to building the *virtual communities* of special interests. It has been observed that blogs not only help in creating new relationships but also enhance existing ones. A seminal analysis on the interpersonal nature of blogs is published in Stefanone and Jang (2008). The blogosphere and social media in general reshapes business models (Scoble and Israel 2006), facilitates viral marketing (Richardson and

Domingos 2002), provides trend analysis and sales prediction capabilities (Gruhl et al. 2005, Mishne and de Rijke 2006), aids counter-terrorism efforts (Coffman and Marcus 2004), and acts as grassroot information sources (Thelwall 2006).

In the physical world, according to Keller and Berry (2003), 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 individuals whose experiences, opinions, and suggestions are sought after are aptly termed as the *influentials* (Keller and Berry 2003). As we draw parallels between physical and virtual communities, among citizens of the blogosphere, we are intrigued by the questions such as whether there exist the influentials in a virtual community (a blog), who they are, and how to find them.

Blogs can be categorized into two major types: *individual* and *community blogs*. 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. In a community blog, however, many have equal opportunities to participate and hence there is a possibility for influential bloggers to emerge. Due to the reason mentioned above, we study community blogs. Henceforth, blogs refer to community blogs.

1.1 Applications of the influentials

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. As reported in (Elkin 2007), approximately 64% advertising companies have acknowledged this phenomenon and are shifting their focus toward blog advertising.

As representatives of communities, the influentials could also *sway* opinions in political campaigns, elections, and affect reactions to government policies (Drezner and Farrell 2004). 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.

According to a report published by Technorati[2] on 5 April 2007, the size of the blogosphere increases by 100% every 6 months. Blogpulse,[3] a blog indexing and tracking website, tracked over 150 K blogs as of 12 December 2010 with over 848 K postings per day. With such a phenomenal growth, novel ways have to be developed in order to keep track of the developments in the blogosphere. Many blog readers/subscribers just want to know the most insightful and authoritative story. Blog posts from the influential bloggers would exactly serve this purpose by standing out as representative articles of a blog site. Being able to identify the influentials is particularly useful to enthusiastic blog readers who often subscribe to several blog sites. The primary focus of this work is to identify such influential bloggers of a community blog that could be considered as representatives by first identifying influential blog posts. This characteristic property of the bloggers is evaluated by studying certain indicators that assist in quantifying community's reactions towards bloggers' postings.[4] However, the influential bloggers identified using the approach proposed in our work could also serve as potential candidates for the applications mentioned above.

1.2 Challenges and contributions

Researchers have studied the influence in the blogosphere from the perspective of influential blog sites (more in Sect. 7) Regardless of a blog being influential or not, it can have its influential bloggers. Influential bloggers of a blog have impact on the fellow bloggers as in a real-world community. In this paper, we address the novel problem of identifying *influential bloggers* in a blog and investigate related issues and challenges.

- Are there influential bloggers as in a real-world community? Are they different from active bloggers?
- What measures should be used to define influential bloggers? A solution can be subjective, depending on the need for identifying influential bloggers.

[1] http://weblogs.macromedia.com/.

[2] http://www.sifry.com/alerts/archives/000436.html.

[3] http://www.blogpulse.com.

[4] More details on identifying and measuring these indicators are provided in Sect. 3.

- How can the influential bloggers be identified? As there is no training data to tell us who are the influential bloggers, it is infeasible to apply supervised 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?

Specifically, we make the following contributions:

- Identify the collectable statistics in the Blogosphere that are used to quantify a blogger's influence.
- Define and formulate the influence of a blogger in terms of the collectable statistics.
- Propose an algorithm—iFinder—that computes the influence score of each blogger.
- Evaluate the proposed algorithm to identify various categories of influential bloggers, their temporal patterns, relative importance of collectable statistics and other interesting observations.
- Design a novel evaluation framework to validate the model in absence of the ground truth.
- Develop a publicly available prototype tool for the proposed model that can be used to crawl, index, and identify influential bloggers in real-world blog sites, besides other analytical capabilities.

In the following, we first define the problem of identifying an influential bloggers in Sect. 2. We then propose a working 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 Sect. 3. Section 4. describes the dataset used in the study. In Sect. 5, we conduct an empirical study to evaluate many aspects of the proposed approach, and observe how the key measures work with a correlation study. We present a publicly available prototype tool for the proposed model that can be used to crawl, index, and identify influential bloggers in real-world blog sites, besides other analytical capabilities in Sect. 6. Section 7 reviews the existing work in this domain. We discuss the potential contribution and significance of our work to social network analysis in Sect. 8. Finally, we conclude our work with future directions in Sect. 9.

2 Influential bloggers: problem and definition

Each blog post is often associated with some metadata like post's author, post annotations, post's date and time, and number of comments. In addition, one can also collect certain statistics from the blog website, e.g., *outlinks* posts or articles to which the author has referred in his/her blog post; *inlinks* other posts that refer to the author's blog post, post length; *average length of comments* per post; and the rate at which comments are posted on a blog post.

In the simplest case, one can approximate an influential blogger with an active blogger who posts frequently. Since in a physical world a voluble person is not necessarily or seldom influential, we are inquisitive whether the same assumption holds in the blogosphere and if we can employ the above metadata and statistics to identify influential bloggers. Hence, the search for influential bloggers boils down to the question as to how to define influence of a blogger. Subsequently, we also need to identify if there are any differences between influential and active bloggers. It is extremely important to identify this difference since it is rather more complex to define an influential blogger leveraging the aforementioned statistics as compared with defining an influential blogger using the activity volume or how frequently a blogger posts. To analyze this distinction, we categorically divide bloggers into four types: active and influential, active and non-influential, inactive and influential, and inactive and non-influential.

Recognizing the subjective nature of influence, we define influential blogger as follows:

Definition 1 *Influential Blogger* A blogger is defined as influential if s/he has at least one influential blog post.

Assume we have an influence score for a post p_i, $I(p_i)$. A blogger could publish several blog posts, some of which could be more influential than others. Based on the definition of an influential blogger, we use the influence score of his/her most influential blog post to determine the blogger's influence. Specifically, for a blogger b_k who has N blog posts, $\{p_1, p_2,\ldots, p_N\}$, their influence scores can be ranked in descending order, and b_k's influence index, iIndex(b_k) can be defined as $\max(I(p_i))$, where $1 \leq i \leq N$. However, there could be a huge variance in the influence scores of the blog posts for some bloggers. In such cases, mean influence score is perhaps a better alternative, which is indicative of a consistent influential blogger. These concepts are defined mathematically in Sect. 3.3. Based on the definition of influential blogger, we can describe the problem statement of identifying influential bloggers as follows:

Problem statement Given a set U of M bloggers, $\{b_1, b_2,\ldots,b_M\}$, the problem of identifying influential bloggers is defined as determining an ordered subset V of K[5] bloggers, $\{b_{j_1}, b_{j_2}, \ldots, b_{j_K}\}$ 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}) $\geq \cdots \geq$ iIndex(b_{j_K}). V contains K most *influential bloggers*. For all the blog posts $\{p_1, p_2,\ldots,p_L\}$ 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

[5] Note that K is a user specified parameter.

$1 \leq l \leq L$. 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}).

We now study the intuitive characteristics that help define iIndex and I, enabling us to build an experimental model that can gauge the influence to distinguish between "influential" and "activeness" properties of bloggers.

3 Identifying the influentials

We first present some desirable properties related to blog post influence which can be approximately defined by collectable statistics, next propose a model for identifying the influentials using these statistics, and then discuss some interesting issues that can be evaluated by experimenting with the model.

3.1 An initial set of intuitive properties

Following Keller and Berry (2003), 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 these influence gestures can be approximated by collectable statistics.

- *Recognition* Social influence depends on the authority that the influential has on the individuals subjected to his/her influence (Turner 1991). The authority or prominence of an actor in directed social networks can be estimated using venerable sociological measures such as prestige and centrality that utilize the edges that are incident upon the actors (Bonacich 1987, Knoke and Burt 1983, Podolny 2005). Similarly, in the blogosphere 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. 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. Recognition of a blog post is measured through the inlinks (ι) to the blog post.
- *Activity Generation* A blog post's capability of generating activity can be indirectly measured by how many comments it receives and 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) indicate 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 (Kolari et al. 2006; Lin et al. 2007).
- *Novelty* Novel ideas exert more influence as suggested in (Keller and Berry 2003). Given the informal nature of the blogosphere, there is no incentive for profuse citations. Based on (Song et al. 2007), outlinks (θ) can be used as an indicator of a post's novelty. If a post refers to many other blog posts or articles it indicates that it is less likely to be novel.
- *Eloquence* An influential is often eloquent (Keller and Berry 2003). 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. Hence, a long post often suggests some necessity of doing so. Therefore, we use the length of a post (λ) as a heuristic to measure eloquence of a blogger. Clearly, length of the blog post is not the best measure to judge the influence of a post since a blogger could ramble on or simply use garbled text. This indicates a need for more sophisticated linguistic measures to examine the writing style. Some measures have been proposed in (Zheng et al. 2006; Argamon et al. 2003) that identify the writing style of articles in online groups using content-based, syntactic, structural, and lexical features. These measures could be used to improve our blog post length based heuristic to determine the eloquence of a blogger. Although a study by Hu et al. in (2007) has reported a positive correlation between length and quality of the articles in Wikipedia, theoretical underpinning with extensive experimental evaluation is left as a possible future research direction to investigate the existence of correlation between blog post length and quality leveraging the research efforts mentioned above in conjunction with the proposed model.

The above four influence gestures form an initial set of properties possessed by an influential post. These four influence gestures with the corresponding statistics collectable from blogs are summarized in Table 1. 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 "hub" blog post. Starting with this initial set, we build a model that allows us to examine, analyze, modify, and extend the model.

Table 1 Influence gestures for identifying influential bloggers and their corresponding collectable statistics

Influence gesture	Collectable statistics	Notation
Recognition	Set of inlinks	ι
Activity generation	Number of comments	γ
Novelty	Set of outlinks	θ
Eloquence	Length of the blog post	λ

3.2 Developing the model

First, we study a model that only uses links to rank the bloggers and then improve on it to include other statistics. For this purpose we consider PageRank (Brin and Page 1998) algorithm that builds upon venerable sociological measures, such as prestige and centrality, to determine an actor's prominence and status in directed social networks, e.g., webpage graph (Bonacich 1987; Knoke and Burt 1983; Podolny 2005). PageRank assigns numerical scores for each blog post, akin to webpages, to "measure" its relative importance as derived from the prominence of the other blogs or webpages from which they receive links or ties. The PageRank score of a blog post (p_i) could also be interpreted as a probability ($R(p_i)$) that represents the likelihood of a random surfer clicking on links, will arrive on this blog post, and is represented as:

$$R(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{R(p_j)}{L(p_j)} \tag{1}$$

where d is the damping factor that the random surfer stops clicking, $M(p_i)$ is the set of all the blog posts that link to p_i, $L(p_j)$ is the total number of outbound links on blog post p_j, and N is the total number of blog posts. The PageRank values $\mathbf{R}$ could be computed as the entries of the dominant eigenvector of the modified adjacency matrix,

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} l(p_1,p_1) & l(p_1,p_2) & \dots & l(p_1,p_N) \\ l(p_2,p_1) & \ddots & & \vdots \\ \vdots & & l(p_i,p_j) & \\ l(p_N,p_1) & \dots & & l(p_N,p_N) \end{bmatrix} \mathbf{R}$$

where the function $l(p_i, p_j)$ is 1 if blog post p_j links to blog post p_i, and 0 otherwise.

As pointed out in (Kritikopoulos et al. 2006), due to the casual environment of the blogosphere, *blog sites are very sparsely linked and it is not suitable to rank blog sites using Web ranking algorithms*. The Random Surfer model of webpage ranking algorithms (Brin and Page 1998) does not work well for sparsely linked network. The sparse adjacency matrix creates challenges in making the adjacency matrix stochastic (explained in more detail in Sect. 3.4), which is a mandatory condition for the convergence of random surfer model to an optimum value. Further, the temporal aspect of blog posts exacerbates the problem of sparsity. A webpage is comparatively a more stable information source. Though the content of the webpage could be dynamic, the URL and the impressions are more static, meaning, over time a webpage is more likely to get recognized and linked by other webpages. On the other hand, blog posts are extremely time sensitive. Each blog post published at a blog site has a unique URL. It is known that on average 18 blog posts are published every second according to Blogpulse statistics in 2010.[6] To put this number in perspective, there were 21.4 million new websites created in 2010,[7] which results in 0.6785 new websites per second. In other words, for every new website that was created in 2010, 26.52 new blog posts appeared on the blogosphere. This demonstrates the extremely dynamic nature of blogosphere as compared with the web. As a consequence of so many new blog posts appearing so frequently, it is extremely challenging to keep track of the original source, thereby making the data on the blogosphere stale too soon. Therefore, while a webpage may acquire links over time, the older a blog post gets the fewer people care about it reducing the chances for the blog post to acquire links over time. Hence, the adjacency matrix of blogs (considered as a graph) will get increasingly sparser as thousands of new sparsely linked blog posts appear every day. The aforementioned differences warrant a novel approach that not only leverages the sparse link graph but also uses other available and relevant statistics to compute influence in the blogosphere. Next, we propose a model—iFinder—that leverages the aforementioned statistics, i.e., inlinks, outlinks, comments, and blog post length. We perform experiments to compare iFinder and PageRank algorithm and report our findings in Sect. 5.

3.3 iFinder: a model to identify influential bloggers

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

[6] http://www.blogpulse.com.

[7] http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/.

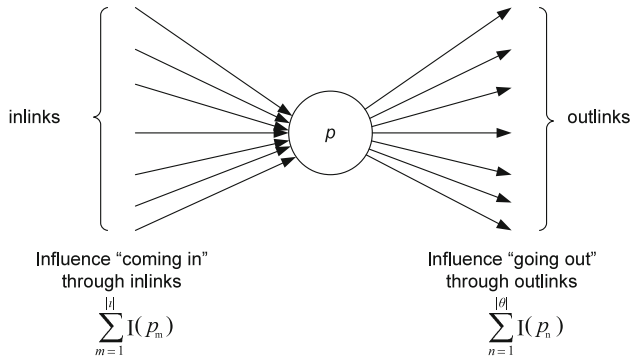


Fig. 1 *i-graph* showing the InfluenceFlow across blog post p

I denotes the influence of a node (or blog post p), then InfluenceFlow across that node is given by,

$$\text{InfluenceFlow}(p) = w_{\text{in}} \sum_{m=1}^{|\iota|} I(p_m) - w_{\text{out}} \sum_{n=1}^{|\theta|} I(p_n) \qquad (2)$$

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. Equation 2 captures the recognition and novelty aspects of the influence gestures. Recognition is estimated by the inlinks a blog post acquires and outlinks account for penalizing the novelty. We illustrate the concept of InfluenceFlow in the i-graph displayed in Fig. 1. This shows an instance of the i-graph with a single blog post. Here we are measuring the InfluenceFlow across blog post p. Towards the right of p are the inlinks and outlinks are towards the left of p. We add up the influence "coming into" p and add up the influence "going out" of p and take the difference of these two quantities to get the influence that p has generated.

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,

$$I(p) \propto w_{\text{com}}\gamma_p + \text{InfluenceFlow}(p) \qquad (3)$$

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 an additive function is good to determine the combined value of each alternative (Fensterer 2007). 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 (Keeney and Raiffa 1993).

From the discussion in Sect. 3.1, 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,[8] 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. 2 and Eq. 3, the influence of a blog post, p, can thus be defined as,

$$I(p) = w(\lambda) \times (w_{\text{com}}\gamma_p + \text{InfluenceFlow}(p)) \qquad (4)$$

The above equation gives an influence score to each blog post. Influence score of each blog post is normalized between 0 and 1. 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. According to the definition of influential blogger in Sect. 2, a blogger can be considered influential if s/he has at least one influential blog post. The influence score of a blogger or iIndex is estimated using the blog post with maximum influence score. There could be other ways to define an influential blogger based on the influence scores of the blog posts published by him/her. For example, if one wants to differentiate a productive or consistent influential blogger from non-prolific one, one might use another measure, such as mean instead of maximum. We can calculate the influence score for each of the blogger B's N posts and use the maximum influence score as the B's iIndex, or,

$$\text{iIndex}(B) = \max(I(p_i)) \qquad (5)$$

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. Blog posts whose influence score is higher than

[8] 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 influence score of the top-kth influential blogger could be termed as *influential blog posts*.

3.4 Computing blogger influence with matrix operations

We have described the iFinder model and how to compute the influence of a blog post using the influence gestures. Here, we convert the computational procedure into basic matrix operations for convenient and efficient implementation.

We define the inlinks and outlinks to the blog posts using a link adjacency matrix $\mathbf{A}$ where the entry A_{ij} is 1 if p_i links to p_j and 0 otherwise, defined as

$$A_{ij} = \begin{cases} 1 & p_i \rightarrow p_j \\ 0 & p_i \nrightarrow p_j \end{cases}$$

Matrix $\mathbf{A}$ denotes the outlinks between the blog posts. Consequently, $\mathbf{A}^T$ denotes the inlinks between the blog posts. We define the vectors for blog post length, comments, influence, and influence flow, respectively, as,

$$\vec{\lambda} = (w(\lambda_{p_1}), \ldots, w(\lambda p_N))^T,$$
$$\vec{\gamma} = (\gamma_{p_1}, \ldots, \gamma_{p_N})^T,$$
$$\vec{i} = (I(p_1), \ldots I(p_N))^T,$$
$$\vec{f} = (f(p_1), \ldots, f(p_N))^T$$

Now, Eq. 2 can be rewritten in terms of the above vectors as,

$$\vec{f} = w_{\text{in}}\mathbf{A}^T\vec{i} - w_{\text{out}}\mathbf{A}\vec{i} = (w_{\text{in}}\mathbf{A}^T - w_{\text{out}}\mathbf{A})\vec{i} \quad (6)$$

and Eq. 4 can be rewritten as,

$$\vec{i} = \text{diag}(\vec{\lambda})(w_c\vec{\gamma} + \vec{f}) \quad (7)$$

Eq. 7 can be rewritten using Eq. 6 which can then be solved iteratively,

$$\vec{i} = \text{diag}(\vec{\lambda})(w_c\vec{\gamma} + (w_{\text{in}}\mathbf{A}^T - w_{\text{out}}\mathbf{A})\vec{i}) \quad (8)$$

or,

$$\vec{i} = (\mathbf{I} - \text{diag}(\vec{\lambda})(w_{\text{in}}\mathbf{A}^T - w_{\text{out}}\mathbf{A}))^{-1}\text{diag}(\vec{\lambda})w_c\vec{\gamma} \quad (9)$$

which is of the form,

$$\vec{i} = (\mathbf{I} - \mathbf{C})^{-1}\mathbf{D} \quad (10)$$

where $\mathbf{C}$ denotes $\text{diag}(\vec{\lambda})(w_{\text{in}}\mathbf{A}^T - w_{\text{out}}\mathbf{A})$ and $\mathbf{D}$ denotes $\text{diag}(\vec{\lambda})w_c\vec{\gamma}$.

The above equation requires $\mathbf{A}$ to be stochastic matrix (Motwani and Raghavan 1995) which means all the blog posts must have at least one outlink. In other words, none of the rows in $\mathbf{A}$ has all the entries as 0. Otherwise, the influence score for such a blog post would be directly proportional to the number of comments. However, in the blogosphere, this assumption does not hold well. Blog posts are sparsely connected. This problem can be fixed by making $\mathbf{A}$ stochastic. This can be achieved by:

- Removing those blog posts with no outlinks and the edges that point to these blog posts while computing influence scores. This does not affect the influence scores of other blog posts, since the blog posts with no outlink do not contribute to the influence score of other blog posts.
- Assigning $1/N$ in all the entries of the rows of such blog posts in $\mathbf{A}$. This implies a dummy edge with uniform probability to all the blog posts from those blog posts which do not have a single outlink.

For a stable solution of Eq. 8, $\mathbf{A}$ must be aperiodic and irreducible (Motwani and Raghavan 1995). A graph is aperiodic if all the paths leading from node i back to i have a length with highest common divisor as 1. One can only link to a blog post which has already been published and even if the blog post is modified later, the original posting date still remains the same. We use this observation to remove cycles in the blog posts by deleting those links that are part of a cycle and point to the blog posts which were posted later than the referring post. This guarantees that there would be no cycles in $\mathbf{A}$, which makes $\mathbf{A}$ aperiodic. A graph is irreducible if there exists a path from any node to any node. Using the second strategy mentioned above by adding dummy edges to make $\mathbf{A}$ stochastic, ensures that $\mathbf{A}$ is also irreducible.

As in (Brin and Page 1998; Kleinberg 1998; Yin et al. 2007), iFinder adopts an iterative method to compute the influence scores of blog posts. iFinder starts with little knowledge and with each iteration tries to improve the knowledge about the influence of the blog posts until it reaches a stable state or a fixed number of iterations specified a priori. The knowledge that iFinder starts with is the initialization of the vector $\vec{i}$. There are several heuristics that could be used to initialize $\vec{i}$. One way to initialize the influence score of all the blog posts is to assign each blog post uniformly a number, such as 0.5. Another way could be to use inlink and outlink counts in a linear combination as the initial values for $\vec{i}$. In our work, we used authority scores from Technorati.[9] One could also use PageRank values to initialize $\vec{i}$, but since we compare our results with PageRank algorithm we do not use it as the initial scores to maintain a fair comparison.

The computation of influence score of blog posts can be done using the well known *power iteration method* (Golub and Van Loan 1996). The underlying algorithm of iFinder

[9] http://technorati.com/developers/api/cosmos.html.

can be described as: Given the set of blog posts P, $\{p_1, p_2,\ldots,p_N\}$, we compute the adjacency matrix $\mathbf{A}$, and vectors $\overrightarrow{\lambda}$ and $\overrightarrow{\gamma}$. The influence vector $\overrightarrow{i}$ is initialized to $\overrightarrow{i_0}$ using Technorati's authority values. Using Eq. 8 and $\overrightarrow{i_0}$, $\overrightarrow{i}$ is computed. At every iteration we use the old value of $\overrightarrow{i}$ to compute the new value $\overrightarrow{i'}$. iFinder stops iterating when a stable state is reached or the user specified iterations are exhausted, whichever is earlier. The stable state is judged by the difference in $\overrightarrow{i}$ and $\overrightarrow{i'}$, measured by cosine similarity. The overall algorithm is presented in Algorithm 1. This algorithms essentially produces the eigenvector with the eigenvalue of 1. As mentioned in (Brin and Page 1998), power iteration method converges in roughly 52 iterations on a database of 322 million links.

Input: Given a set of blog posts P, number of iterations *iter*, Similarity threshold τ
Output: The influence vector, $\overrightarrow{i}$ which represents the influence scores of all the blog posts in P.

Compute the adjacency matrix $\mathbf{A}$;
Compute vectors $\overrightarrow{\lambda}$, $\overrightarrow{\gamma}$;
Initialize $\overrightarrow{i} \leftarrow \overrightarrow{i_0}$;
repeat
  $\overrightarrow{i'} = \overrightarrow{\lambda}(w_c\overrightarrow{\gamma} + (w_{in}\mathbf{A}^T - w_{out}\mathbf{A})\overrightarrow{i})$;
  $iter \leftarrow iter - 1$;
until $(cosine\_similarity(\overrightarrow{i}, \overrightarrow{i'}) < \tau) \vee (iter \geq 0)$;

3.5 Issues of identifying the influentials

The proposed model presents a tractable 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 iFinder (cf. Sect. 5.1).
- How can we evaluate iFinder's performance in identifying the influential bloggers? Are influential blog posts different from non-influential blog posts? (cf. Sect. 5.2).
- Since there is no training and test data, how do we evaluate the efficacy of the proposed model? 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 (cf. Sect. 5.3).
- How does iFinder perform when compared against other models to find authoritative blog posts like PageRank (Brin and Page 1998)? (cf. Sect. 5.4).
- 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 can these weights help find special influential bloggers? (cf. Sect. 5.5).
- Are all the four parameters necessary? Is there a correlation between the parameters, making some of them redundant? How can the model be extended? Are there any other parameters that can be incorporated in a refined model? (cf. Sect. 5.6).
- 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? (cf. Sect. 5.7).

In the next sections, we set out to use the proposed model in an empirical study, seek answers to the aforementioned questions, analyze results, report findings, and suggest new lines of research in finding influential bloggers.

4 Data collection

Here, we discuss the need for experimental data and select a real-world blog site for experiments. Data collection is one of the critical tasks in this work. 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[10] 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, The Unofficial Apple Weblog (TUAW),[11] containing the statistics required by iFinder. The advantages of of doing so include,

1. Minimizing our effort in figuring out ways to obtain the needed statistics, and
2. Maximizing the reproducibility of our experiments independently.

[10] http://www.nielsenbuzzmetrics.com/cgm.asp.
[11] http://www.tuaw.com/.

TUAW provides 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 is obtained using Technorati API.[12] We crawled the TUAW blog site and retrieved over 10,000 blog posts published between February 2004[13] and 31 January 2007. We keep the complete history of the TUAW blog site and update it incrementally. All the statistics obtained after crawling are stored in a relational database for fast retrieval later.[14]

Table 2 Two lists of the top five bloggers according to TUAW (most active) and iFinder (most influential)

Five most active TUAW bloggers	Five most influential bloggers using iFinder
Erica Sadun	*Erica Sadun*
Scott McNulty	Dan Lurie
Mat Lu	*David Chartier*
David Chartier	*Scott McNulty*
Michael Rose	Laurie A. Duncan

5 Experiments and further study

Next, we design various experiments with the proposed model using iIndex and answer the questions raised in Sect. 3.5 based on the experimental results. In the process, we develop and elaborate an evaluation procedure for effective comparison.

5.1 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 study window of 30 days as well. The number of posts of a blogger 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 our first experiment, we generate a list of k most influential bloggers using the model proposed in Sect. 3.3. 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 Sect. 5.2.2. By setting $k = 5$, we compare the five most active bloggers published at TUAW with five most influential bloggers obtained using iFinder in Table 2, where the first column contains the five most active bloggers published by TUAW and the second column lists the five most influential bloggers. Names in *italics* are the bloggers present in both lists. Three out of five TUAW most active bloggers are also among the top five most influential bloggers identified by iFinder. 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' are in the TUAW list, so they are active; and 'Dan Lurie' and 'Laurie A. Duncan' are 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 3.

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 3.

- *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 nine 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 four of his blog posts are influential. One of his influential posts is about iPhone,[15] which attracted a large number of bloggers to comment and triggered a heated discussion of the new product (77 comments and 33 inlinks). Its length is 1,417 bytes, and there are no outlinks. All

[12] http://technorati.com/developers/api/cosmos.html.

[13] TUAW was setup in February 2004.

[14] This dataset will be made available upon request for research purposes.

[15] http://www.tuaw.com/2007/01/09/iphone-will-not-allow-user-installable-applications/.

Table 3 Comparison of statistics between different bloggers

		Number of comments		Number of inlinks		Blog post length		Number of outlinks		Total number of blog posts	Influential blog posts
		Max	Avg	Max	Avg	Max	Avg	Max	Avg		
Active & influential	Erica Sadun	75	11.02	80	10.13	2,935	830	15	2.53	152	9
	David Chartier	56	11.31	32	10.25	3,529	1,055	14	4.35	68	4
	Scott McNulty	112	11.56	33	8.925	2,246	623	12	2.59	107	3
Inactive & influential	Dan Lurie	96	19.63	37	10.26	1,569	794	4	2.32	16	4
	Laurie Duncan	65	16.29	34	10.61	2,888	994	11	3.47	26	2
Active & non-influential	Mat Lu	42	8.029	29	10.01	1,699	771	12	4.1	73	0
	Michael Rose	31	8.727	21	9.606	1,378	736	15	6.15	58	0

Table 4 Comparison of statistics between influential and non-influential blog posts

	Number of comments		Number of inlinks		Blog post length		Num of outlinks		Total number of blog posts
	Max	Avg	Max	Avg	Max	Avg	Max	Avg	
Influential blog posts	112	74.18	80	38.63	3,529	1,999.32	15	3.36	22
Non-influential blog posts	69	10.84	39	8.96	1,930	703.74	27	4.3	513

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. These kind of posts allows a reader to experience something new and 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 five most influentials. 'Mat Lu' and 'Michael Rose' were ranked 3rd and 4th 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 five most influential bloggers because their other statistics are not comparable with those of the influentials (i.e., having fewer comments and inlinks, and more outlinks).

5.2 Influential versus 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 Sect. 2, we pick influential blog posts submitted by the influential bloggers listed in Table 2. 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 3, 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 4. It shows influential blog posts are much longer in length, have far more comments, and attract lot more inlinks. Although influential blog posts have fewer outlinks, it is not a very strong distinguishing feature as compared to inlinks, comments, and blog post length due to smaller difference margin. A more detailed analysis on the parameters is presented in Sect. 5.6.

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,[16] 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[17] with a developer.[18] This incident triggered a lot of discussion through 57 comments and 20 inlinks. Many people commented and cited this blog post, and agreed

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

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

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

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.

5.3 Evaluating the model

The absence of ground truth presents a huge challenge to evaluate the efficacy of the proposed model. 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.[19] 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 and thus will not appear in Digg. For January 2007, there were in total 535 blog posts submitted on TUAW. As Digg only returns the 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. For categories S1, S2, and S3, we rank the blog posts based on the influence score and pick top 20 blog posts from each of the three categories. We randomly pick 20 blog posts from the category S4, where 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 5. From the table, we can see that 85% of the 20 most influential blog posts published by bloggers that belong to S1 (i.e., influential and active set of bloggers) make it to Digg's list of top 100. The results show the differences among the four categories of bloggers and iFinder 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 6 and 7, respectively. Note that we selected top five active and five influential bloggers (Table 2), in which three are both active and influential (Table 3). We observe from Tables 5, 6 and 7 that influential bloggers have higher chances to be liked than active bloggers. We explain this observation in the following analysis:

1. Compare active and influential bloggers (S1) with active and non-influential bloggers (S3) in Tables 6 and 7. 21.71% (= 71/327) of blog posts from S1 were liked by people (judged by their votes on Digg), but only 6.1% (= 8/131) of blog posts from S3 were liked by the people, according to Digg. This shows that the chances of being liked by the people are more if the blogger is influential and not if he/she is active. Results from iFinder in Table 5 are also consistent with this observation. Furthermore, compare influential and active bloggers (S1) with influential and inactive bloggers (S2) in Tables 6 and 7. 21.71% (= 71/327) of blog posts from S1 were liked by the people and 33.33% (= 14/42) of the blog posts from S2 were liked by the people. This shows that regardless of the blogger being active or inactive, if he is influential he is liked more by the people (judged by the votes on Digg). Results from iFinder in Table 5 are also consistent with this observation. These two facts bring out the difference between influential and active bloggers. Influential bloggers are more liked as compared with active bloggers.
2. According to S3 in Tables 5, 6, and 7, active bloggers are not necessarily influential while according to S1, influential bloggers may be active.
3. In Table 6, S4 has seven blog posts liked by people even though they were non-influential and inactive. This is because one of the bloggers in S4 was ranked sixth in the list of influential bloggers and four of his blog posts appeared in Digg. So in such cases where the blogger is on borderline we could get good overlap values for S4 too.

Table 5 Percentage of 20 most influential blog posts published by each of the four different categories of bloggers at TUAW that appeared on Digg

Bloggers	Active	Inactive
Influential	S1: 85% (=17/20)	S2: 35% (=7/20)
Non-influential	S3: 15% (=3/20)	S4: 5% (=1/20)

iFinder was used to identify the most influential blog posts for each category of bloggers in TUAW

[19] http://www.digg.com/.

5.4 iFinder versus pagerank

We compared iFinder with Google Pagerank. We used Google's blog search interface to obtain the ranked list of blog posts according to the PageRank values because of two primary reasons. First, we do not implement the PageRank algorithm to avoid concerns regarding accurate implemenation. Second, Google keeps on evolving their algorithm so the search interface has the latest and most advanced version of the PageRank which is certainly better than the primitive and published version of the PageRank (Brin and Page 1998).

We compared the 20 most influential blog posts in a pairwise fashion from iFinder, PageRank, and Digg for each month starting from January 2007 to June 2007. The results are reported in Table 8. Unlike Digg there is no issue of coverage with Google's PageRank comparison. Google indexes all the blog posts available at TUAW. Since our comparison is on monthly basis, we check this by looking at the total number of results displayed by Google and the total number of blog posts submitted at TUAW for each month. It is evident from Table 8 that first, iFinder performs better than Google's PageRank when compared with Digg as the ground truth. Second, to rule out the possible explanation that Digg does not cover all the blog posts so Google-Digg overlap is poor; we study the overlap between Google and iFinder, since both cover all the blog posts submitted. Results show that there is an insignificant overlap between Google and iFinder. Third, Google's model is less aligned with Digg as compared with iFinder, which shows that Google's blog post relevance ranking does not fit well with the taste of the people.

Table 6 Distribution of 100 TUAW's blog posts that appeared on Digg grouped by the bloggers belonging to one of the four categories

Bloggers	Active	Inactive
Influential	S1: 71% (=71/100)	S2: 14% (=14/100)
Non-influential	S3: 8% (=8/100)	S4: 7% (=7/100)

Note that Digg's API only returns 100 blog posts ranked in decreasing number of votes, or also known as "diggs"

Table 7 Distribution of the 535 TUAW blog posts grouped by the bloggers belonging to one of the four categories

Bloggers	Active	Inactive
Influential	S1: 61.112% (=327/535)	S2: 7.850% (=42/535)
Non-influential	S3: 24.485% (=131/535)	S4: 6.542% (=35/535)

5.5 Effects and usages of weights

There are four weights in the proposed model to regulate the contribution of four parameters toward the calculation of the influence score using Eqs. 2 and 4. 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 normalization purposes (outside the scope of this work).

For the remaining three weights, w_{comm}, w_{in} and w_{out}, we fix two and observe how the ranking changes by varying the third. 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 $w_{\text{comm}} \geq 0.6$, 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 $w_{\text{in}} \geq 0.9$ and $w_{\text{out}} \geq 0.2$. To summarize, we obtain the same ranking of influential bloggers as shown in the right column of Table 2 for $w_{\text{comm}} \geq 0.6$, $w_{\text{in}} \geq 0.9$, $w_{\text{out}} \geq 0.2$.

Clearly, changing the value of the above three weights can lead to different rankings. This allows one to adjust the weights of the model to identify 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. Larger value for w_{out} can be set to discourage the citations of other blog

Table 8 Overlap between iFinder, Google PageRank, and Digg (20 most influential blog posts from each model)

	Jan-07 (%)	Feb-07 (%)	Mar-07 (%)	Apr-07 (%)	May-07 (%)	Jun-07 (%)
iFinder and Digg	60	50	75	60	80	70
PageRank and Digg	20	15	30	20	40	20
PageRank and iFinder	30	20	30	25	35	25

posts encouraging a post with independent ideas. If one wants to emphasize one aspect, one can tune weights and obtain ranking to reflect that aspect. In short, these weights provide a means to further evolve and expand the proposed model for a wide range of applications.

5.6 Parameter study

We conduct more experiments to: (1) verify if any of the four factors (number of comments, inlinks, outlinks, and length of a blog post) can be eliminated via a lesion study; (2) examine the pairwise correlations of the four factors; (3) observe the relative relevance of all the parameters; and (4) conduct more experiments to study another statistic—*the rate of comments* to extend iFinder.

5.6.1 Lesion study

We study the performance of the model by removing one parameter in turn. That is, we compute the influence scores using *only* the remaining three parameters. We rank the five most influential bloggers by leaving one parameter out and thus obtain four ranking results, comparing with the result of "All-in" (with all four parameters). Had there been a parameter that did not contribute to the influence score, removing it would not result in any difference in the ranking. The results are presented in Fig. 2. The x-axis denotes different ranking schemes to find the influentials. For example, "No outlinks" signifies the ranking of influential bloggers computed using inlinks, comments, and post length, but leaving outlinks out. Interestingly, all the top five influentials remain unchanged, but their relative ranks vary. It is evident that no blogger maintains the same rank in all the five variations and no two ranked lists are the same. Thus, the four parameters contribute in the proposed model in determining influential blogger. As discussed in Sect. 5.5, the trade-off between the parameters can be achieved by adjusting their associated weights to accommodate different needs.

5.6.2 Correlation analysis

We perform pairwise correlation analysis between the parameters to further examine whether there is any redundant parameter. With four parameters, there are six pairwise correlations as shown in Fig. 3a–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 is substitutable. We notice that five of six 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

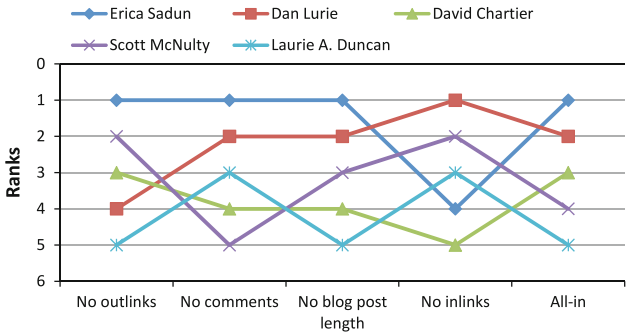


Fig. 2 Evaluating significance of each of the parameters through lesion study

receives, and vice versa. This supports that links among blog posts are different from web links (Sect. 7).

5.6.3 Relative relevance of parameters

Since Digg assigns score to blog posts and not bloggers, we compare the top most influential blog posts from Digg[20] and iFinder. We compare 20 most influential blog posts[21] for every month for the last 6 months starting from January 2007 till June 2007. We report the overlap in the two lists. Since there is not 100% overlap, rank correlation coefficients like Kendall-Tau rank correlation coefficient (Kendall 1938) or Spearman's rank correlation coefficient (Spearman 1904) could not be computed. We try different configurations of iFinder by considering,

1. All-in, i.e., all the four parameters,
2. No inlinks (outlinks, comments, and blog post length),
3. No comments (inlinks, outlinks, blog post length),
4. No outlinks (inlinks, comments, blog post length), and
5. No blog post length (inlinks, outlinks, comments).

We report the overlap results for all these five configurations with Digg in Table 9.

We get the best overlap for "All-in" configuration for May 2007, i.e., 80 and 50% overlap for February 2007. For "All-in" configuration, on average we achieve 65.83% overlap. Although the proposed model is not at par in terms of overlap with the users' opinion on Digg, which indicates a possibility for improvement, it provides an objective strategy to evaluate a subjective concept.

We also studied the contribution of different parameters and their relative importance from the experiments with the other four configurations. From the results in Table 9, it can be observed that configuration 2 (no inlinks) always performs the worst, configuration 3 (no comments)

[20] We get this data using Digg API.

[21] On average, 70–80 blog posts from TUAW are submitted to Digg every month, so we pick 20 most "digged" or influential posts to avoid under-sampling or over-sampling.

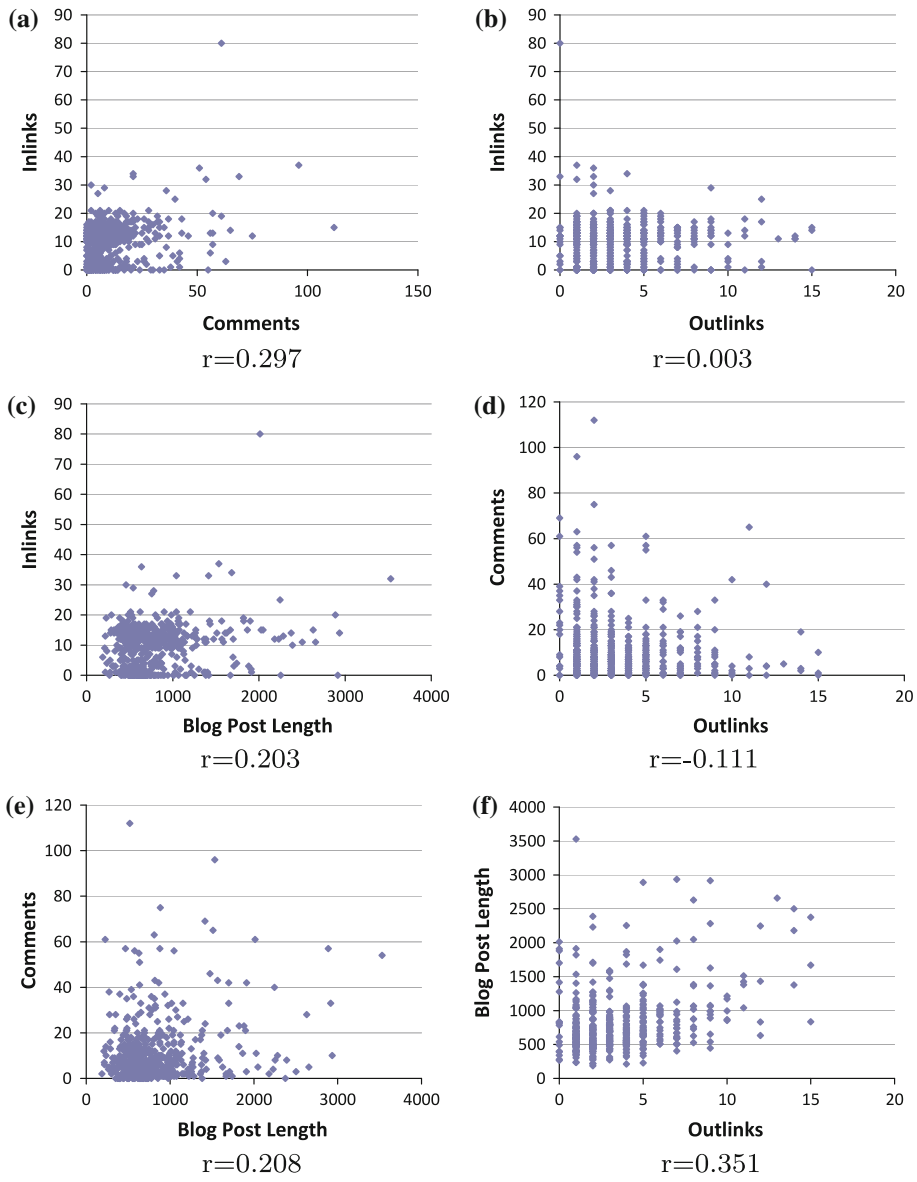


Fig. 3 Pairwise correlation plots of the four parameters (ι, θ, λ, and γ) of the blog posts

performs better, then comes configuration 4 (no outlinks), and then come configuration 5 (no blog post length). This gives us the order of importance of all the four parameters, i.e., inlinks > comments > outlinks > blog post length, in the decreasing order of importance to influence estimation. Given this analysis, we can adjust the weights for different parameters to achieve better than "All-in" results.

5.6.4 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 Figs. 4 and 5 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. Fig. 4 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 Fig. 5 are relatively "flat", around 10 comments per hour even after 7 or 8 h 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

Table 9 Overlap between 20 most influential blog posts at Digg and iFinder for last 6 months for different configurations

	Jun-07 (%)	May-07 (%)	Apr-07 (%)	Mar-07 (%)	Feb-07 (%)	Jan-07 (%)
All-in	70	80	60	75	50	60
No inlinks	15	20	15	15	5	0
No comments	40	40	20	16	20	16
No outlinks	55	40	20	16	16	35
No blog post length	60	70	55	75	45	50

special time-critical applications like disaster prevention and management, emergency handling.

Other extensions to the proposed model include the following:

1. Study of spam comments filtering to prevent spam attacks using techniques mentioned in (Kolari et al. 2006; Lin et al. 2007),
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 high penalty for large 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 $w_{\text{out}} \sum_{n=1}^{|\theta|} I(p_n)$ in Eq. 2 with $\exp(w_{\text{out}} \sum_{n=1}^{|\theta|} I(p_n))$. We would have to investigate thoroughly the role of w_{out} in such a scenario.

5.7 Temporal patterns of the influentials

In the above experiments we studied 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 in February 2004. We hence apply iFinder to identify top five 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 February 2004–January 2007. The temporal patterns of the influentials can be observed from a matrix in Fig. 6. 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 jth blogger in the ith time window. For example, the first cell (*sean bonner*, *Feb-04*) shows that *Sean Bonner* was ranked first among

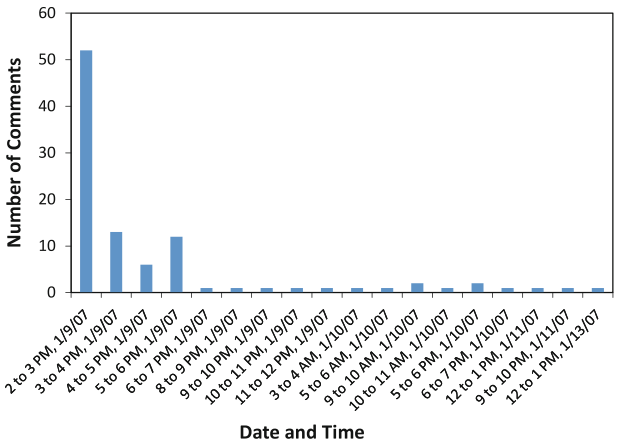


Fig. 4 Spiky comments reaction on a blog post related to iPhone

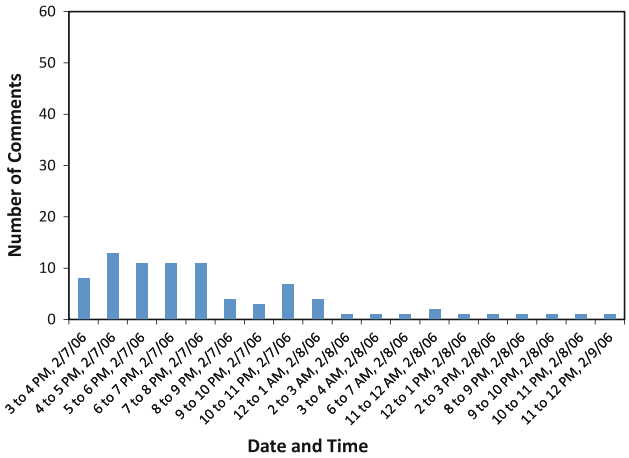


Fig. 5 "Flat" comments reaction on a blog post related to some competition in Apple Inc

the influential bloggers list in February 2004.[22] Gray cells represent that the particular blogger was not among the five most influentials for that time period. The color gradient represents rank of a influential blogger, a darker color representing a higher rank.

We can observe some different temporal patterns for the influentials in Fig. 6. 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

[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 five influentials.

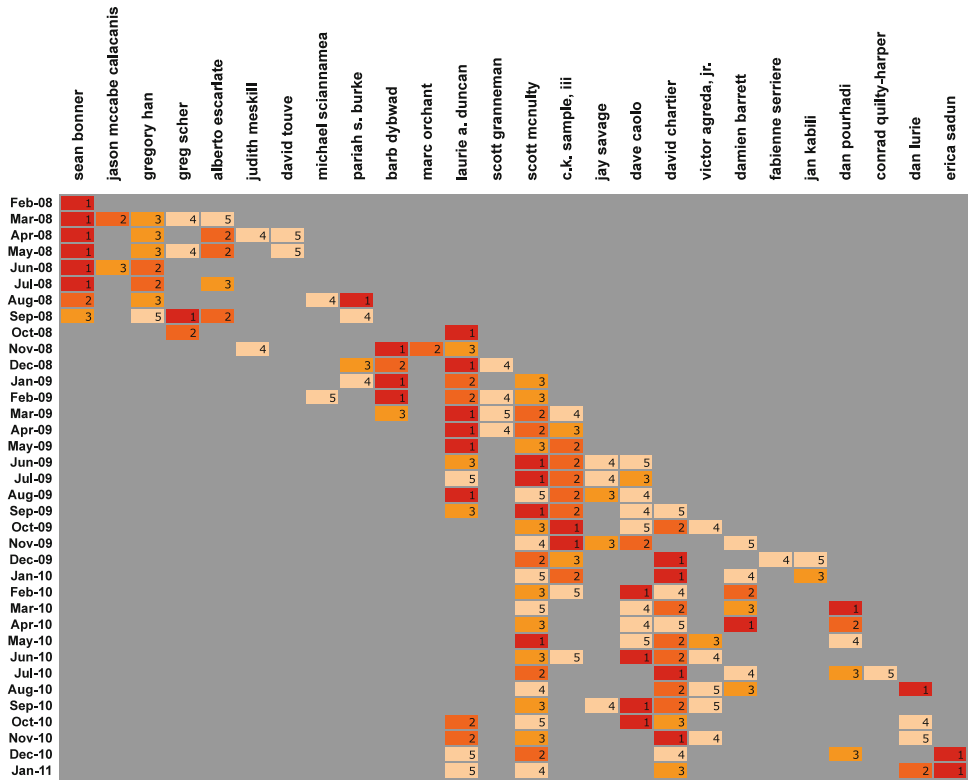


Fig. 6 Influential Bloggers' blogging behavior over the whole TUAW blog history. The number in the cells indicate the influence rank of the bloggers (1 being the most influential and 5 being the least influential). A colored version of the figure is available at http://ualr.edu/nxagarwal/iFinder/TemporalPattern.pdf

Fig. 6 are "Sean Bonner", "Gregory Han", and "Barb Dybward".

Transient influentials They are influential for a very short time period (only 1 or 2 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 of more follow-up examinations.

As observed above, bloggers could exhibit different temporal patterns. Many potential applications can be developed using these patterns. Long-term influentials are more reliable as compared with other bloggers due to a successful history of being recognized as influential for a prolonged period of time. When we want to know about a blog site, the best way to approach it is to look at its long-term influentials as they have lasting influence in the community. Blog posts of the 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 influential blogger. Further analysis is needed to investigate the role of longitudinal patterns of influential bloggers in the aforementioned applications. The categories presented here are some examples. Certainly, there could be several other temporal patterns and can find uses in other applications.

5.8 Experiment on engadget

In the previous sections, we evaluated iFinder on TUAW dataset. We performed a similar study on another blog site, Engadget.[23] Specifically, we analyzed iFinder's generalizability towards other blogs. We studied whether iFinder can effectively identify influential bloggers on other blog sites. We evaluated iFinder on Engadget data using Digg as a reference point. We also compared the results with Google's PageRank through their blog search interface. We collected same statistics from Engadget such as, inlinks, outlinks, number of comments, blog posts, and other metadata such as date and time of blog post, blogger name, and permalink of the blog post. We presented a prototype tool called BlogTrackers in Sect. 6 that could be used to crawl the blog site and extract these information pieces and store it in relational database.

We conducted experiment on Engadget data for the period of 6 months starting from January 2009 to June 2009. We identified influential blog posts on a monthly basis for this period. The reason for selecting this time window was due to the API restrictions of Digg. Digg allows fetching the data from their database for a window

[23] http://www.engadget.com.

of 30 days. This is similar to the previous evaluation on TUAW data. In order to construct a reference ranking list, we collected all the blog posts of Engadget that were submitted to Digg for the period of January 2009 to June 2009, sliced into 30-day time windows. Total number of such blog posts for every time window is denoted by D, as shown in the second column of Table 10. We ranked these blog posts in the decreasing order of their digg scores. For each month, we computed n most influential blog posts using iFinder and PageRank. We varied n between 10 and 100. We computed "precision at n" statistic for both the models, iFinder and PageRank. The statistics, precision at n, or $P@n$, is computed as

$$P@n = \frac{|\{\text{rel}\} \cap \{\text{ret}\}|}{|\{\text{ret}\}|} = \frac{|\{\text{rel}\} \cap \{\text{ret}\}|}{n} \tag{11}$$

where n is the total number of documents retrieved, {rel} denotes the set of relevant documents, |{rel}| denotes the total number of relevant documents, {ret} denotes the set of retrieved documents, and |{ret}|denotes the total number of retrieved documents. $P@n$ is same as precision except that it evaluates the ranking algorithms at different cut-off ranks, considering only the top results returned by the ranking algorithm. The results are shown in Table 10. n is varied from 10, 20, 30, 50, 100, and D. Average $P@n$ statistics in Fig. 7 show that iFinder consistently reports higher precision at different values of n as compared with PageRank. This shows that iFinder can be used to identify influential bloggers on different blogs as long as the required statistics are available. Value for $P@n$ decreases as n increases because from Eq. 11 it can be observed that n, which is the total number of retrieved documents (|{ret}|), is the denominator of the fraction. Since we have a limited set of relevant blog posts denoted by {rel}, if we increase number of retrieved blog posts, then the precision would reduce. A decrease in the value of $P@n$ can be observed as n increases for both iFinder and PageRank, although the decrease is faster for iFinder than for PageRank. After an in-depth analysis it was observed that iFinder had identified relevant blog posts quite early (for smaller values of n) as compared with PageRank. This led to a faster decrease in $P@n$ value as n (the number of retrieved documents) was increased.

6 BlogTrackers: a prototype tool for iFinder

In previous sections, we discussed the proposed model, illustrated its efficacy, demonstrated its capability to identify various trends, patterns, and categories of the influentials. Inspired by the needs and interests of social scientists and their ways of studying subjects in social

Table 10 Precision@n (P@n) statistics for iFinder and PageRank using Digg as a reference point for the Engadget Data

Month	Top blog posts present on Digg (D)	Total blog posts on Engadget (T)	Precision@n (P@n)											
			$n = 10$		$n = D$		$n = 20$		$n = 30$		$n = 50$		$n = 100$	
			iFinder	PageRank	iFinder	PageRank	iFinder	PageRank	iFinder	PageRank	iFinder	PageRank	iFinder	PageRank
Jan-09	29	1,209	0.7	0.2	0.66	0.24	0.7	0.3	0.63	0.23	0.56	0.22	0.29	0.21
Feb-09	11	1,020	0.6	0.4	0.64	0.36	0.5	0.3	0.37	0.20	0.22	0.16	0.11	0.11
Mar-09	25	993	0.7	0.5	0.64	0.24	0.7	0.3	0.60	0.20	0.5	0.2	0.25	0.18
Apr-09	18	1,068	0.8	0.3	0.78	0.28	0.8	0.3	0.60	0.30	0.36	0.28	0.18	0.18
May-09	22	1,011	0.8	0.2	0.73	0.41	0.8	0.4	0.73	0.43	0.44	0.32	0.22	0.2
Jun-09	20	985	0.5	0.2	0.60	0.30	0.6	0.3	0.57	0.33	0.4	0.32	0.2	0.2
	Average P@n		0.68	0.30	0.67	0.31	0.67	0.31	0.58	0.28	0.41	0.25	0.21	0.18

D denotes the total number of Engadget blog posts found on Digg for a particular month

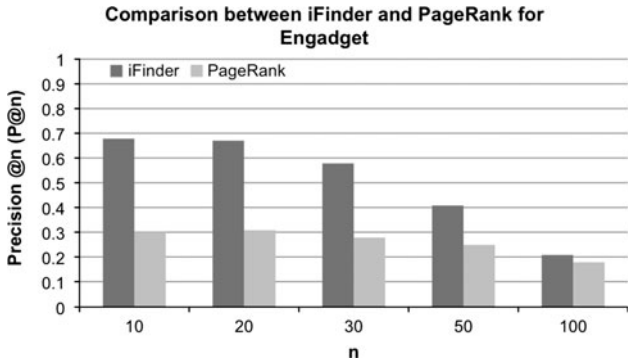


Fig. 7 Comparison between iFinder and PageRank on Engadget for P@n values averaged over 6 months (January 2009 to June 2009)

media, we built a prototype called, BlogTrackers[24] that provides a generic platform to collect data in the blogosphere, identify key topics of discussion, track blogs of particular interests over time, identify influential bloggers, facilitate comparative data analysis, and search blogs. These features can help social scientists to quickly analyze both blogs and bloggers at a scale, which is otherwise impossible through manual investigation. In this section, we briefly describe the features and functionalities of BlogTrackers.

BlogTrackers (Agarwal et al. 2009a, 2009b) is a Java-based desktop application that provides a unified platform for the user to crawl and analyze blog data. BlogTrackers supports the analysis of one or more blog sites simultaneously. The BlogTrackers system is composed of two modules, namely, the crawler and the tracker. The crawler module is responsible for retrieving blog data from the blogosphere, indexing, and storing it in a relational database. We collect data and metadata from blog sites, such as the comments received by a blog post, inlinks of a blog post, outlinks of a blog post, the blogger name, the timestamp of a blog post, various categories and tags associated with a blog post, and actual blog post content. The tracker module of BlogTrackers uses these individual pieces of information to support analysis capabilities, like identification of influential bloggers, topic detection, and tracking future developments of a topic with the help of keywords.

BlogTrackers offers two kinds of crawlers to retrieve blog data, batch crawler and RSS crawler. The batch crawler uses the screen scraping techniques to crawl blog sites and retrieve archived blog posts. The blogosphere is comprised of many popular blogging platforms like Blogger, WordPress, and BlogSpot. One of the biggest hurdles in crawling the blogosphere is the lack of API support by the blog sites to retrieve the data from their blog posts. This forces us to use techniques like screen scraping to crawl the blogs. This method is not very scalable, especially when the blogs are highly customized. To understand the complexity in creating a generic blog crawler, we have built customized crawlers for 50 Indonesian blogs to gauge the ability to write a generic crawler. Most of these blogs used a customized version of the popular WordPress blogging platform and we were able to successfully retrieve information, such as the number of comments, tags, categories, and the blog post.

Bloggers use Really Simple Syndicate (RSS) feeds to automatically inform their subscribers about new blog posts. This format is well structured and frequently used by the bloggers. Our RSS crawler can periodically retrieve the latest blog posts from the blog sites using the RSS feeds of the individual blog sites. The crawler is fault tolerant and the user has complete control of its scheduling options. As mentioned earlier, the blog data sources we currently track include the two popular technology blog sites: The Unofficial Apple Weblog (TUAW), Engadget, and 50 popular Indonesian blogs.

BlogTrackers uses the iFinder model proposed in this article to identify influential bloggers during a certain time interval. The interface is very intuitive and allows the users to vary the weight associated with each of the parameters and generate influential scores for bloggers in the system. The interface is shown in Fig. 8a, with the ranking of the bloggers from TUAW during the period 24–27th January 2010. In Fig. 8b, we can observe the blog posts associated with the most influential blogger. We can also observe the keywords corresponding to the blogger during this period. In this case, the most frequent keywords for the blogger consist of "iPad" and "iPhone".

The users also have the flexibility to correlate the activity with the influence of bloggers. The tool can classify the bloggers into four different categories: "Active-Influential", "Inactive-Influential", "Active-Non Influential", and "Inactive-Non Influential" based on their influence scores and their activity during a specific time interval.

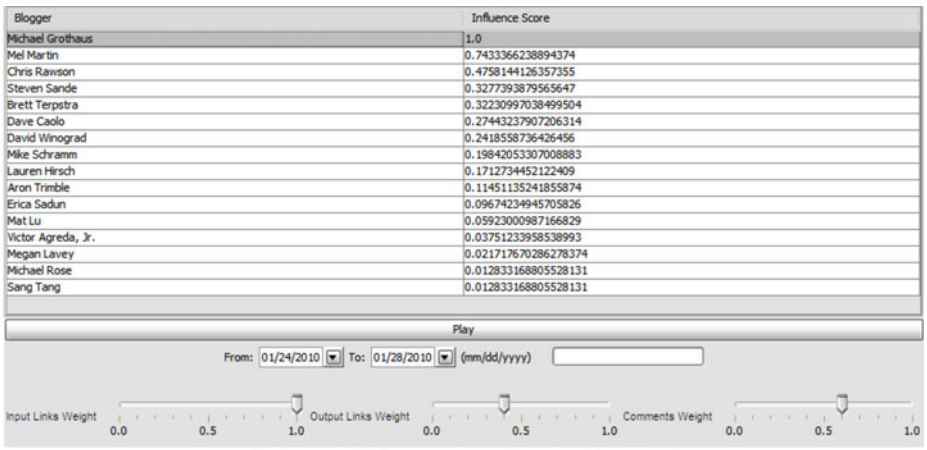
7 Related work

The blogosphere has been expanding speedily since its inception. This has attracted a surge of research on the blogosphere. Authors in (Chin and Chignell 2006) 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 overlap between different sub-communities are used to help identify influence between them. Next, we review briefly existing works in the area of influential blog sites and bog leaders. We compare and contrast these approaches with the work proposed in this paper.

[24] http://blogtrackers.fulton.asu.edu/.

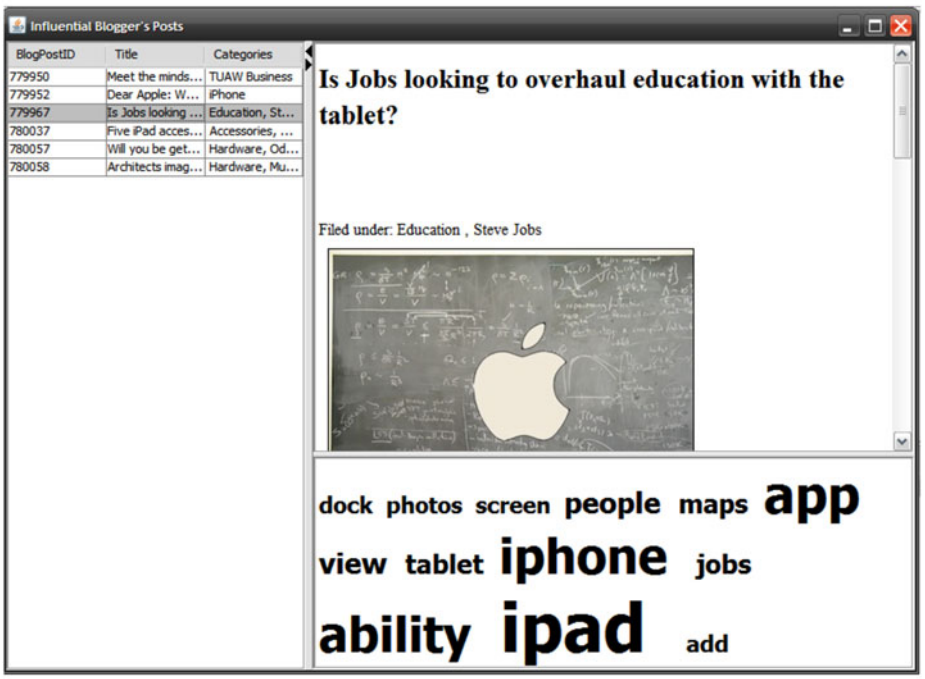
Fig. 8 Influential blogger analysis in BlogTrackers

(a)



Influential Blogger interface in BlogTrackers

(b)



Blog posts and top keywords for the top influential blogger

7.1 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 (Gill 2004). It is different from the problem of identifying influential bloggers in a community. The blogosphere follows a power law distribution (Faloutsos et al. 1999), with very few influential blog sites forming the short head of the distribution and a large number of non-influential sites forming the long tail where abundant new business, marketing, and development opportunities can be explored (Anderson 2006). Our work identifies influential bloggers at a blog site regardless of the site being influential or not. We briefly review some work on identifying influential blog sites.

Researchers have studied blog graph from the perspective of information diffusion and identify key players who maximize the spread (Leskovec et al. 2007). Gruhl et al. (Gruhl et al. 2004) study information diffusion of various topics in the blogosphere, drawing on the theory of infectious diseases. A general cascade model (Goldenberg et al. 2001) 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 (Richardson and Domingos 2002; Kempe et al. 2003; Chen et al. 2009) 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) (Java et al. 2006). Leskovec et al. (Leskovec et al. 2007) proposed a submodularity-based approach to identify the most important blogs which outperforms the greedy approach. Nakajima et al. (Nakajima et al. 2005) attempts to find *agitators*, who stimulate discussions; and *summarizers*, who summarize discussions, by thread detection. The phenomenal growth of the blogosphere along with increasing link sparsity presents significant challenges to employ purely link analysis based approaches.

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), and comments which are 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.

Interestingly, Watts and Dodds (2007), Watts (2007) challenged the conventional "influential hypothesis". They conducted computer simulations of interpersonal influence processes on networks of different generation models and found that large cascades of influence are driven by a critical mass of easily influenced individuals rather than a few influentials. Hence, they proposed, first, to generate a sufficiently large number of seed adopters and then, rely on the viral property of interpersonal influence to reach more actors (Watts and Peretti 2007). *iFinder* can be used to find a local influential, which may be useful for cost-effective seeding for communities in the long tail.

7.2 Blog leaders

In the past, researchers have studied various forms of blog leaders in the blogosphere. Here we analyze these different types of blog leaders and compare with influential bloggers. Authors in (Ni et al. 2007) categorize the blogs into two classes, "Affective Blogs" and "Informative Blogs". Affective blogs are personal accounts and experiences. Informative blogs are more technology oriented, news related, commonsense knowledge, objective, and high-quality informative blogs. Training a binary classifier on a hand-labeled set of blogs using Naïve Bayes, SVM, and Rocchio classifier they separate the affective blogs from the informative blogs. However, there could be influential bloggers who write affective blogs which would be missed by such an approach.

Several works such as (Kavanaugh et al. 2006; Song et al. 2007) study the existence of "Opinion Leaders" in the physical world and observe their role when they engage in some form of blogging. In the physical world, opinion leadership is reflected in the degree to which an individual is able to informally influence other individuals in the form of shaping or changing their attitudes or overt behavior in a desired way with relative frequency (Rogers 1995; Rogers and Shoemaker 1971; Merton 1968; Katz 1957; Coleman et al. 1966; Berelson et al. 1986; Katz and Lazarsfeld 1955; Lazarsfeld et al. 1944). Accordingly, opinion leadership is earned and maintained by individual's technical competence, social accessibility, and conformity to the social system's norms. Similarly, influential blog sites in the blogosphere exert influence over the external world and within the blogosphere (Gill 2004). Song et al. (2007) define opinion leaders on the blogosphere as those who bring in new information, ideas, and opinions, then disseminate them down to the masses. They rank the blogs using a novelty score measured by the difference in the content of a given blog post and ones that refer to this blog post. First, the blog posts are reduced to topic space using Latent Dirichlet allocation (LDA), and then, using cosine similarity measure between these transformed blog posts novelty score is computed. However, opinion leaders could be different from influential bloggers. There could be a blogger who is not very novel in his/her content but attracts a lot of attention to his/her posts through comments and feedback. These type of bloggers will not be captured by simply novelty based approaches. Moreover, due to the casual nature of the blogosphere, not many blogs refer to the original source they borrowed their content from, making it challenging to determine novelty of a blog and hence the opinion leaders.

When user action history is available, it is possible to learn the influence probabilities between users (Goyal et al. 2010). The authors shows that the influence probabilities between users can be learned based on user action history of joining groups in Flickr. However, collecting user action history in the blogosphere is easier said than done. For instance, the behavior of audience cannot be crawled unless they leave traces such as comments. The blog host might be able to save user browsing information, but this can seldom be shared with the third-parties.

Many blog sites list "Active 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) (Gill 2004). Certainly these statistics would leave out those blog sites or bloggers who were not active. Moreover, influential bloggers are not necessarily active bloggers at a blog site.

7.3 Bibliometrics

Identifying influential articles is an important area of study under bibliometrics. Some existing work has performed citation analysis (Moed 2005) on bibliometrics datasets such as DBLP.[25] Here we highlight some key differences between influential articles in citation datasets and identifying influentials in the blogosphere. First, authors cite the articles they refer in their papers which is not the case in the blogosphere. Due to casual nature of blogs, bloggers seldom link to the articles/blog posts they were inspired from. This creates an extremely *sparse* network structure unlike the dense citation network. Second, blog data have additional information such as comments which is not available in bibliometrics data. Such information is useful in estimating a blog post's influence in the community. Due to this difference, studies in bibliometrics focus on network structure derived by author co-citation analysis (Chen and Paul 2001), while research proposed in the blogosphere can include both network structure properties as well as content statistics as illustrated by the proposed model.

It is noted that some recent work attempts to identify topic-sensitive influential authors (Mimno and McCallum 2007; Tang et al. 2009) by joint analysis of contents, authorship, and citations. The topics are typically extracted from content information (i.e., the paper contents), and the influentials are based on corresponding link analysis. A similar idea is also applied to the case of social media such as Twitter (Weng et al. 2010). In this work, however, we aim to address the problem of finding influential bloggers within a community, who rather focus on a specific topic. For example, in our experiments, the community studied are concentrating on Apple or broadly technology-related information. We expect that in the blogosphere, this kind of topic modeling based on tags and blog posts can also be applied to help find topic-sensitive influential bloggers.

8 Significance to social network analysis

The blogosphere represents an interconnection of blogs that exist together as a connected community or a collection of connected communities. The network of blogs implicitly defines a social network of bloggers where bloggers can share their personal experiences, opinions, and views, communicate with fellow bloggers by expressing concerns via comments, and connect with other blogs or bloggers via links. A blogger network is a special case of online social networks that provides information not only about blogger ties but also about the blogger themselves and what other bloggers think about them.

In this work, we developed a model to identify influential bloggers in a community. The influential bloggers could be considered as the prominent actors in the blogosphere who are authoritative and possess the capability to affect people's opinions, choices, and attitudes (Rogers 1995; Rogers and Shoemaker 1971; Merton 1968; Katz 1957; Coleman et al. 1966; Berelson et al. 1986; Katz and Lazarsfeld 1955; Lazarsfeld et al. 1944). This could be compared with existing sociological measures that determine an actor's prominence in a social network using centrality and prestige (Bonacich 1987; Knoke and Burt 1983; Podolny 2005) metrics. However, the existing measures completely rely on the network information to identify prominent actors, which could be underestimated due to casual nature of the blogosphere (Kritikopoulos et al. 2006). Due to the sparse network information, existing measures need to be modified to address this challenge and leverage additional attributes available in the blogosphere. The model proposed here attempts to advance the existing social network analysis measures for the online social media by incorporating the challenges and opportunities that come along with the social media. Specifically, the existing social network analysis measures could be advanced by leveraging the influence gestures mentioned in this work such as comments and blog post quality in addition to the inlinks and outlinks. For other forms of social media such as YouTube, Flickr, Delicious, etc., various additional statistics such as number of hits/views, comments, and subscriptions could lend deeper insights into identifying influential individuals or influential content by complementing the information available through links. The model enables us in differentiating between various types of prominent actors in the blogosphere such as active and/or influentials and identify various longitudinal patterns. Further, we developed an objective evaluation methodology for subjective concepts like influence that not only facilitate verification of the proposed model but can also be used to evaluate other models to identify influential nodes in online social networks. This evaluation methodology eliminates the need for the tedious and expensive process of human subject evaluation, bridging the gap between qualitative and quantitative evaluation measures.

We envisage that the methodology developed in this work can benefit conventional sociological measures. It can be extended to apply to online social media and incorporate various additional attributes. These extensions could also lead to advanced computational models to better our understanding of conventional sociological theories, assist in developing new ones, and reinforcing the development of more accurate and efficient social interaction modeling algorithms for diverse environments. We anticipate that the

[25] http://kdl.cs.umass.edu/data/dblp/dblp-info.html.

lessons learned from this research would create greater synergies between social science and computational science.

9 Conclusions and future work

We address a novel problem of identifying influential bloggers in a community blog site. Our work differs from existing works on identifying influence in traditional media, identifying influential blog sites in the blogosphere, and influence maximization within the blogosphere. Influential bloggers can exist in any blog site, regardless of the site 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 social media site, and conduct experiments by using the whole history of blog posts of real-world blog sites. The results demonstrate that

1. active bloggers are not necessarily influential bloggers,
2. by tuning the weights associated with the parameters of the proposed model, one can examine how different parameters impact the influence ranking for different needs,
3. the proposed model can serve as a baseline in identifying influential bloggers and can be extended by not only incorporating additional parameters to discover different patterns but also sophisticating existing parameters such as length of the blog post by including linguistic analysis, and
4. Categorizing the influentials into different groups paves the way for applying many existing data mining algorithms to discover deeper patterns.

We expect that the proposed model will evolve to address many new needs arising from the real (or rather virtual) world.

In this work, we proposed a model to identify influential bloggers in a community blog site. However, the blogosphere consists of more individual blogs than community blogs. Therefore, in future we would like to extend the study to include individual blogs too. However, since individual blogs are single authored, it is inappropriate to find influential bloggers. This could be achieved by synthesizing a virtual community of similar individual blogs. We have published some works that aggregate individual blogs that are similar and were not connected previously (Agarwal et al. 2007, 2009c). These aggregated individual blogs could be treated as a virtual community. However, different blogs have different outreach. This makes some blogs more visible and hence are read more, which requires normalization of the model parameters proposed in our study and exploration of new parameters.

Acknowledgments This research was funded in part by the National Science Foundations Social-Computational Systems (SoCS) Program within the Directorate for Computer and Information Science and Engineerings Division of Information and Intelligent Systems (Award numbers: IIS-1110868 and IIS-1110649), the US Office of Naval Research (Grant number: N000141010091), and the US Air Force Office of Scientific Research (Grant number: FA95500810132). We gratefully acknowledge this support.

References

Agarwal N, Kumar S, Lim M, Liu H (2009a) Mapping socio-cultural dynamics in indonesian blogosphere. In: 3rd AAAI International Conference on Computational Cultural Dynamics (ICCCD09), pp 37–44

Agarwal N, Kumar S, Liu H, Woodward M (2009b) Blogtrackers: a tool for sociologists to track and analyze blogosphere. In: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM)

Agarwal N, Liu H, Murthy S, Sen A, Wang X (2009c) A social identity approach to identify familiar strangers in a social network. In: Proceedings of the Third International AAAI Conference of Weblogs and Social Media, pp 2–9

Agarwal N, Liu H, Salerno JJ, Yu PS (2007) Searching for familiar strangers on blogosphere: problems and challenges. In: NGDM

Anderson C (2006) The long tail: why the future of business is selling less of more. Hyperion, New York

Argamon S, Koppel M, Fine J, Shimoni A (2003) Gender, genre, and writing style in formal written texts. TextInterdiscip J Study Discourse 23(3):321–346

Berelson B, Lazarsfeld P, McPhee W (1986) Voting: a study of opinion formation in a presidential campaign. University of Chicago Press, Chicago

Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92(5):1170–1182

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 30(1–7):107–117

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. In: Proceedings of the seventh international conference on World Wide Web, pp 107–117

Chen C, Paul R (2001) Visualizing a knowledge domain's intellectual structure. Computer 34(3):65–71

Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, pp 199–208

Chin A, Chignell M (2006) A social hypertext model for finding community in blogs. In: HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia, ACM Press, New York, pp 11–22

Coffman T, Marcus S (2004) Dynamic classification of groups through social network analysis and hmms. In: Proceedings of IEEE Aerospace Conference

Coleman J, Katz E, Menzel H (1966) Medical innovation: a diffusion study. Bobbs-Merrill Co, Indiana

Drezner D, Farrell H (2004) The power and politics of blogs. In: American Political Science Association Annual Conference

Elkin T (2007) Just an online minute… online forecast. http://publications.mediapost.com/index.cfm?fuseaction=Articles.showArticle&artaid=29803

Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. In: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication, pp 251–262

Fensterer GD (2007) Planning and assessing stability operations: a proposed value focus thinking approach. PhD thesis, Air Force Institute of Technology

Gill KE (2004) How can we measure the influence of the blogosphere? In: Proceedings of the WWW'04: workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics

Gillmor D (2006) We the media: grassroots journalism by the people, for the people. O'Reilly, Sebastopol

Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Mark Lett 12:211–223

Golub G, Van Loan C (1996) Matrix computations. 3rd edn. Johns Hopkins University Press, Baltimore

Goyal A, Bonchi F, Lakshamanan LVS (2010) Learning influence probabilities in social networks. In: WSDM

Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. In: KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM Press, New York, pp 78–87

Gruhl D, Liben-Nowell D, Guha R, Tomkins A (2004) Information diffusion through blogspace. SIGKDD Explor Newsl 6(2):43–52

Hu M, Lim E, Sun A, Lauw H, Vuong B (2007) Measuring article quality in wikipedia: models and evaluation. In: Proceedings of the Sixteenth ACM Conference on Conference on information and Knowledge Management, ACM, New York, pp 243–252

Java A, Kolari P, Finin T, Oates T (2006) Modeling the spread of influence on the blogosphere. In: Proceedings of the 15th International World Wide Web Conference

Katz E (1957) The two-step flow of communication: an up-to-date report on an hypothesis. Public Opin Q 21(1):61–78

Katz E, Lazarsfeld P (1955) Personal influence: the part played by people in the flow of mass communications. Free Press, Glencoe, IL

Kavanaugh A, Zin TT, Carroll JM, Schmitz J, Manuel Pérez-Qui N, Isenhour P (2006) When opinion leaders blog: new forms of citizen interaction. In: Proceedings of the 2006 international conference on Digital government research, ACM, New York, pp 79–88

Keeney RL, Raiffa H (1993) Decisions with multiple objectives: preferences and value tradeoffs. Cambridge University Press, Cambridge

Keller E, Berry J (2003) 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, New York

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the KDD, ACM Press, New York, pp 137–146

Kendall M (1938) A new measure of rank correlation. Biometrika 30:81–89

Kleinberg J (1998) Authoritative sources in a hyperlinked environment. In: 9th ACM-SIAM Symposium on Discrete Algorithms

Knoke D, Burt R (1983) Prominence. In: Applied network analysis, pp 195–222

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

Kritikopoulos A, Sideri M, Varlamis I (2006) 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, ACM Press, New York

Lazarsfeld P, Berelson B, Gaudet H (1944) The People's Choice. How the Voter Makes up His Mind in a Presidential Campaign 1944. Columbia University Press, New York

Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, pp 420–429

Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M (2007) Cascading behavior in large blog graphs. In: SIAM International Conference on Data Mining

Lin Y-R, Sundaram H, Chi Y, Tatemura J, Tseng BL (2007) 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), ACM press, New York, pp 1–8

Merton R (1968) Social theory and social structure. Free Press, New York

Mimno D, McCallum A (2007) Mining a digital library for influential authors. In: JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, ACM, New York, pp 105–106

Mishne G, de Rijke M (2006) 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, ACM Press, New York, pp 925–926

Moed H (2005) Citation analysis in research evaluation. Kluwer Academic Publishers, Dordrecht

Motwani R, Raghavan P (1995) Randomized algorithms. Cambridge University Press, Cambridge

Nakajima S, Tatemura J, Hino Y, Hara Y, Tanaka K (2005) Discovering important bloggers based on analyzing blog threads. In: Annual Workshop on the Weblogging Ecosystem

Ni X, Xue G-R, Ling X, Yu Y, Yang Q (2007) Exploring in the weblog space by detecting informative and affective articles. In: WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM, New York, pp 281–290

O'Reilly T (2005) 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

Podolny J (2005) Status signals: a sociological study of market competition. Princeton University Press, Princeton

Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data mining, ACM Press, New York, pp 61–70

Rogers E (1995) Diffusion of innovations. Free Press, New York

Rogers E, Shoemaker F (1971) Communication of innovations: a cross-cultural approach. Free Press, New York

Scoble R, Israel S (2006) Naked conversations: how blogs are changing the way businesses talk with customers. Wiley, London

Song X, Chi Y, Hino K, Tseng B (2007) Identifying opinion leaders in the blogosphere. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, New York, pp 971–974

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

Stefanone M, Jang C (2008) Writing for friends and family: the interpersonal nature of blogs. J ComputMediat Commun 13(1):123–140

Tang J, Sun J, Wang C, Yang Z (2009) Social influence analysis in large-scale networks. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, pp 807–816

Thelwall M (2006) Bloggers under the London attacks: top information sources and topics. In: Proceedings of the 3rd annual workshop on webloging ecosystem: aggreation, analysis and dynamics

Turner J (1991) Social influence. Thomson Brooks/Cole, Belmont

Watts D (2007) Challenging the influentials hypothesis. WOMMA Meas Word Mouth 3:201–211

Watts D, Dodds P (2007) Influentials, networks, and public opinion formation. J Consum Res 34(4):441

Watts DJ, Peretti J (2007) Viral marketing in the real world. Harvard Business Review, Cambridge

Weng J, Peng Lim E, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential twitterers. In: WSDM

Yin X, Han J, Yu PS (2007) Truth discovery with multiple conflicting information providers on the web. In: IEEE Transactions on Knowledge and Data Engineering (TKDE)

Zheng R, Li J, Chen H, Huang Z (2006) A framework for authorship identification of online messages: writing-style features and classification techniques. J Am Soc Inf Sci Technol 57(3):378–393