Identifying the Productive and Influential Bloggers in a Community

Leonidas Akritidis, Dimitrios Katsaros, and Panayiotis Bozanis

***Abstract*—Social networking has become one of the most important trends on the Web, leading to the development of several social applications such as blogs. Blogs are locations on the Web where individuals are provided with the ability to express their opinion, experience, and knowledge about a product, an event, or any other subject. The tremendous popularity of these services has rendered the problem of identifying the most influential bloggers significant, since its solution can lead to numerous major benefits for commerce, advertising, and searching. The current works on this topic either ignore temporal aspects or they fail to gracefully incorporate recency, productivity, and influence at the same time. This paper investigates the issue of identifying bloggers who are both productive and influential by introducing the blogger's productivity index and blogger's influence index. The proposed metrics are evaluated against the state-of-the-art influential blogger identification methods by employing data collected from a real-world community blog site. The obtained results confirm that the new methods are able to identify significant patterns in the bloggers' behavior.**

***Index Terms*—Blogosphere, influential bloggers, ranking.**

I. Introduction

The dynamic and participatory features that were introduced by the deployment of Web 2.0 led to the development of several novel Web services with interactive characteristics. One of the most important of these features allowed the users to communicate in environments that are currently known as social networks. Blogs are a significant form of social networking; they are locations on the Web where individuals publish and share thoughts, experiences, and opinions about various subjects such as social events, products, services, news, and so forth. The visitors of these services are provided with the ability to access the content of a blog, submit comments, or even express their judgments by voting.

Blogs have rapidly gained popularity and attracted the attention of the users, since they provide high functionality, and publication is quite straightforward. Some reasons that made blogging so popular stem from the human psychology [1]: Some people use blogs to inform others about their activity and whereabouts, to document their lives, to provide commentary and opinions, to express deeply felt emotions, and many others. According to the most recent report [2] of the Technorati blog search engine,[1] there are about 184 million bloggers world wide who publish roughly 1 million posts on a daily basis.

In a real-world community, people tend to consult others when they are about to make a decision. Such decisions include purchases, event attendances, travel destinations, or even political voting. Similarly, the blogosphere is a virtual world where the users ask and listen to the opinions of the bloggers on various aspects of life, such as which restaurant to choose, which place to visit, or which movie to watch. Hence, they are influenced by others in their decision making and these others are defined in [3] as the influentials.

The influentials are usually connected in large virtual communities. Consequently, they play a special role in multiple ways. For instance, commercial companies can turn their interest in gaining the respect of the influentials to become their "unofficial spokesmen" instead of investing huge amounts of money and time to advertise their products to thousands of potential customers. Some statistics extracted from the aforementioned Technorati report [2] reveal that a significant percentage of bloggers negotiated directly with advertisers (approximately 20%), whereas 33% of them used affiliate advertising on their blog sites. Apart from their commercial and advertising significance, the influentials could also be responsible for forging political agendas by affecting the voting behavior of their readers.

The recent explosion of the blogosphere has attracted a surge of research on issues related to it [4]. The identification of influential blog sites [5] and the related study of the spread of influence among blog sites [3], [6], [7], [8] are orthogonal to the problem considered in this paper, which focuses on identifying influential bloggers in a single blog site. Finally, it is obvious that works proposing methodologies for discovering and analyzing blog communities [9], [10] are not relevant to the present problem.

The problem of identifying the influential bloggers has been introduced in [11] where a specific model is presented to confront it. That initial model—*the influence flow method*—explicitly discriminated the influential from the active (i.e., productive) bloggers, and considered features specific to the blogosphere, like the size of the blog post, the number of comments, and the incoming and outgoing links. Nevertheless, the blogosphere is a rapidly changing landscape and this model fails to incorporate temporal aspects which are crucial to the influential blogger discovery problem. Furthermore, this model does not take into account the productivity as a factor which affects the influence. On the other hand, the metric for identifying a blogger's influence (*MEIBI*) and *MEIBI* extended *(MEIBIX)* metrics presented in [12] take into consideration all these factors and they support time-aware identification of the influentials. However, these metrics assign a unique value to each blogger and they cannot provide a straightforward separation between influence and productivity.

Motivated by the weaknesses of the existing approaches, this paper proposes a new way of identifying influential and at the same time productive bloggers in community blogs. It proposes time-aware metrics by considering both the temporal and productivity aspects of the blogging behavior, along with the interlinkage among the blogs posts. Apart from the aforementioned benefits that derive from the identification of influential bloggers, the proposed metrics could be exploited by blog or Web search engines in order to effectively rank blog-oriented pages, provide valuable assistance in the recognition of the experts in a particular subject area at a specific time instance, or to be used for providing recommendations [13].

The rest of the paper is organized as follows. Section II introduces the proposed algorithms for the identification of the influential bloggers. Section III evaluates the proposed algorithms with a dataset obtained from a real-world blog community, and finally, Section IV concludes this paper.

II. Identifying Currently Productive and Influential Bloggers

In the following, we identify the factors that play a crucial role in the measurement of a blogger's influence and how these factors can be expressed as mathematical equations.

Manuscript received May 25, 2010; revised September 17, 2010; accepted December 1, 2010. Date of publication January 13, 2011; date of current version August 19, 2011. This paper was recommended by Associate Editor M. Last.

The authors are with the Department of Computer and Communication Engineering, University of Thessaly, Volos 38221, Greece (e-mail: leoakr@inf.uth.gr; dkatsar@inf.uth.gr; pbozanis@inf.uth.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCC.2010.2099216

[1]http://www.technorati.com.

A. Factors Measuring a Blogger's Influence

Beyond any doubt, the number of incoming links to a blog post is strong evidence of its influence. Similarly, the number of comments made to a post is another strong indication that this blog post has received significant attention by the community. The case of outlinks is more subtle.In Web ranking algorithms like PageRank and hyperlink-induced topic search (HITS), the links are used only as a recognition of (or to convey) authority. The influence-flow method of [11] assigns two semantics to a link: it is the means of conveying authority, and it is also a means of reducing the novelty. This mechanism results in two significant problems: 1) It misinterprets the intention of the link creators; and 2) it causes stability and convergence problems to the algorithm for the influence score calculation due to the existence of two sums (one for inlinks and one for outlinks). It is characteristic that the authors admit [11, page 215] that the presence of outlinks in novel posts is quite common and it is used "to support the post's explanations." Therefore, we argue that the outlinks are not relevant to the post's novelty, and all links should have a single semantic, that of implying endorsement (influence).

It is generally acceptable that the longer documents are possibly of higher informational value than the shorter ones. This intuition is also present in some of the most successful Web ranking functions, such as BM25 [14], where the length of a document is a factor that determines its score during query processing. Regarding blog communities, although the length of a post is not a safe indication of its influence, we accept that longer posts are likely to cause stronger reactions from other bloggers or readers than the shorter ones.

The temporal dimension is of crucial importance for identifying the influentials in a rapidly changing environment such as blogosphere. Time is related to the age of a blog post and also to the age of the incoming links to that post. Moreover, the age of the comments made to a post is also of significant importance. In the former case, the time involves the age of the post (e.g., in days since the current day) and in the latter case, the time involves the age (e.g., in days since the current day) of the incoming links and of the comments.

An influential blogger is recognized as such if he/she has written influential posts recently or if the posts have had an impact recently. Specifically, the impact can be as follows.

1) *Proximal impact:* Denotes the influence that a blogger has on the regular members/readers of the community. It is mainly visible by the comments made to a post.
2) *Wide impact:* Denotes the influence that a blogger has on other bloggers outside the community. The incoming links that a post receives is a strong indication of this type of impact. There are also some other indications revealing the impact that a post has outside the community (such as the number of Facebook or Twitter shares), but these characteristics are supported only by a limited number of blog communities and are not considered in this paper.

There is another observation evident by the analysis presented in [11]: Many of the influential bloggers were also active (i.e., productive) bloggers [11, Tabs. 1 and 3–5]. Although productivity and influence do not coincide, there is a strong relation between them.

B. Blogger Productivity and Influence

In this section, we examine how we can evaluate the productivity and the influence of a blogger with respect to the unique characteristics of the blogosphere. The methods proposed here satisfy the requirements described in Section II-A, and the formulae used to estimate the productivity and the influence of a blogger are developed in a way that keeps computational requirements at low levels, without concerning the stability and convergence issues encountered in the implementation of the influence-flow model. Some useful notation is presented in Table I.

TABLE I
NOTATION

Symbol	Meaning
P_t^j	the productivity of blogger j at time t
I_t^j	the influence of blogger j at time t
N^j	the set of blog posts of blogger j
n_i^j	i-th blog post of blogger j
C_i^j	the set of comments to post i of blogger j
R_i^j	set of posts referring (have link to) the i-th post of blogger j
L_i^j	the length (in words) of the i-th post of blogger j
$\overline{L}$	the average post length (in words)
t	time variable
$t_{i,p}^j$	timestamp of the post i of blogger j
$t_{x,l}^j$	timestamp of the inlink x referring to a post of blogger j
$t_{x,c}^j$	timestamp of the comment x submitted to a post of blogger j

As already mentioned, the blogosphere changes rapidly in a manner that a blogger who was considered as productive or influential at a certain point in time may not remain productive or influential in the future. New bloggers enter the community whereas others leave it and thousands of posts are submitted every day. In Section III, it is demonstrated that a blogger may submit up to hundreds (or even thousands) of posts yearly. In this dynamic environment, the time that a blogger's post was submitted is crucial, since a blog post becomes "old" very quickly. A remarkable topic, which is now of major importance, may be totally outdated after two months. Similarly, the submission date of the incoming links and the comments is also significant, since it reveals in general how influential a blogger is presently.

To account for this, it is assigned a time-varying score $U_{i,p}^j(t)$ to the ith post of the jth blogger as follows:

$$U_{i,p}^j(t) = \gamma \frac{L_i^j}{\overline{L}} \left(\frac{\theta}{t - t_{i,p}^j + \theta} \right)^{\delta} \tag{1}$$

where γ, δ, and θ are predefined constants. The parameter γ is not absolutely necessary, but it is used to grant to the quantities $U_{i,p}^j(t)$ a value large enough to be meaningful. The quantity $t - t_{i,p}^j$ represents the time between the current date and the publication date of the post expressed in seconds.[2] If a time difference expressed in days is required, then an appropriate value for θ can be chosen (i.e., $\theta = 86\,400$). Similarly, the parameter δ does not affect the relative score values in a crucial way, but it is used to determine the rate at which the older posts decay. Therefore, by modifying its value, we can determine how quickly a post becomes outdated. Both parameters do not need complicated tuning, since they are not absolutely necessary; in our experiments, γ and δ are assigned values equal to 100 and 1, respectively.

The definition of scores $U_{i,p}^j(t)$ embodies the relation between the value of a post and its age; the scores decay over time. Based on this definition, a new metric is introduced, namely *blogger's productivity (BP) index*, for evaluating the productivity of an individual blogger. The definition of *BP*-index is as follows:

Definition 1 (BP-index): In a given time instance t, a blogger j has *BP*-index equal to P_t^j, if P_t^j of his/her N^j posts get a score $U_{i,p}^j(t) \geq P_t^j$ each, and the rest $N^j - P_t^j$ posts get a score of $U_{i,p}^j(t) < P_t^j$.

[2]The timestamp we use in this paper is merely a 32-bit integer indicating the number of seconds between a particular date and January 1, 1970.

This definition awards the productivity of a blogger and according to it, a blogger will be currently productive if he/she has posted several long posts recently.

In the following, we examine how the influence of a blogger can be calculated. As we have already mentioned, the influence of a blog post has a dual nature, and it consists of the proximal impact (revealed by the comments it receives) and the wide impact, which is expressed by the number and the age of the incoming links. If a post is not cited or no longer commented, it is an indication that it negotiates outdated topics or proposes outdated solutions. On the other hand, if an old post continues to be linked until presently, then this is an indication that it contains influential material. Equation (2), shown below, reflects this dual nature by assigning to each blog post a score determined by both the comments and the inlinks.

$$V_{i,p}^{j}(t) = w_l \sum_{\forall x \in R_i^j} \left(\frac{\theta}{t - t_{x,l}^{j} + \theta}\right)^{\delta} + w_c \sum_{\forall x \in C_i^j} \left(\frac{\theta}{t - t_{x,c}^{j} + \theta}\right)^{\delta}. \quad (2)$$

The parameters w_l and w_c function similarly to γ; that is, they are used to grant the score $V_{i,p}^{j}(t)$ a reasonably large value. However, one may argue that a comment is not as valuable as an incoming link, and these parameters can be used to regulate the desired balance. In our experiments, we have used the combination, $w_l = 100$ and $w_c = 10$, which means that each incoming link is considered as important as ten user comments.

Based on (2), the definition of the *Blogger's Influence (BI) Index* metric is formulated as follows:

Definition 2 (BI -index): In a given time instance t, a blogger j has *BI*-index equal to I_t^j if I_t^j of his/her N^j posts get a score $V_{i,p}^{j}(t) \geq I_t^j$ each, and the rest $N^j - I_t^j$ posts get a score of $V_{i,p}^{j}(t) < I_t^j$.

This definition rewards the bloggers whose posts are receiving many comments and incoming links presently and it addresses the issue of identifying bloggers with recent influence.

The introduction of this family of metrics generates a straightforward policy for evaluating the temporal productivity and influence of the bloggers. No user-defined weights need to be set before these metrics provide results, whereas the most sound features of the blogosphere are considered. Moreover, the calculation of the metrics can be performed in an online fashion, since they do not involve complex computation and they do not present stability problems like those encountered when using eigenvector-based influence scores. Note that the developed metrics are similar in spirit with the h-index and its variations [15] that recently became popular in the scientometrics literature, but the challenges in the blogosphere are completely different: there are comments associated with each blog post, the time granularity is finer, the author of a post is a single person, the resulting graph might contain cycles, and much more.

At this point, it is useful to recall two metrics proposed earlier in [12] which are inspired by the work reported in [16]. The first one, namely *MEIBI*, assigns a score $S_j^m(i)$ to the ith post of the jth blogger as follows:

$$S_j^m(i) = c_1(C_i^j + 1)R_i^j \left(\frac{\theta}{t - t_{i,p}^{j} + \theta}\right)^{\delta} \quad (3)$$

where $c_1 = 4$, $\theta = 86\,400$, and $\delta = 1$. Using the definition of scores $S_j^m(i)$, the definition of *MEIBI* is phrased accordingly.

Definition 3 (MEIBI-index [12]): A blogger j has *MEIBI* equal to m if m of his/her N^j posts get a score $S_j^m(i) \geq m$ each, and the rest $N^j - m$ posts get a score of $S_j^m(i) < m$.

TABLE II
DATASET CHARACTERISTICS

	Engadget
Bloggers	93
Posts	63,358
Inlinks	319,880
Comments	3,672,819
Comments per Post	57.97
Inlinks per Post	5.05
Posts per Author	681.27
Average Post Length (words)	180.52

TABLE III
INCOMING LINKS AND COMMENT AGE WITH RESPECT TO THE PUBLICATION DATE OF THE ORIGINAL POST

Age	**Inlinks**	**Comments**
0 days	132,204 (41.3%)	2,693,745 (73.3%)
1 day	40,360 (12.6%)	476,896 (13.0%)
between 2 and 7 days	39,470 (12.3%)	208,834 (5.7%)
between 8 and 30 days	26,866 (8.4%)	46,169 (1.3%)
between 31 and 60 days	14,535 (4.5%)	26,192 (0.7%)
between 61 and 365 days	43,320 (13.5%)	171,291 (4.7%)
over 365 days	23,125 (7.2%)	49,692 (1.4%)
Total	319,880	3,672,819

The second metric, i.e., *MEIBIX*, operates similarly to *MEIBI*, but instead of assigning to a blogger's old posts smaller scores depending on their age, it assigns to each incoming link of a blogger's post a smaller weight depending on the link's age. This idea is quantified into the following equation:

$$S_j^x(i) = c_1(C_i^j + 1) \sum_{\forall x \in R_i^j} \left(\frac{\theta}{t - t_{x,l}^{j} + \theta}\right)^{\delta} \quad (4)$$

Based on (4), the definition of the metric is formulated similarly as follows:

Definition 4 (MEIBIX-index [12]): A blogger j has *MEIBIX*index equal to x if x of his/her N^j posts get a score $S_j^x(i) \geq x$ each, and the rest $N^j - x$ posts get a score of $S_j^x(i) < x$.

III. EVALUATION

In this section, we evaluate our metrics against the influence flow model: *MEIBI* and *MEIBIX* . For the evaluation, we need active blog communities that provide all the necessary data characteristics, such as the publication date of a post, its corresponding author, and the number of comments that each post received along with their submission date. The Engadget (http://www.engadget.com) technology blog is a community meeting all these requirements.

Table II records the most important characteristics and the sizes of our obtained dataset. The final row indicates the average post length in words, including the title. To obtain the total number and publication date of the incoming links, we employed the blog search service of Google (http://blogsearch.google.com/) by requesting the posts including references to the posts of our dataset. Apart from the number of the incoming links, we also retrieved the date that the referring post was submitted and the name of its corresponding author.

Table III depicts the time distribution of all inlinks and comments made to the original posts of Engadget. We observe that during the first week of its life, a post receives the 66.2% of its total references and the 92% of its comments. This is an indication that time-aware analysis is

TABLE IV
BLOGGERS RANKING BASED ON THE h-INDEX

	Bloggers	h	Posts (Cited)	Inlinks	Cmnts
1	R. Block	53	5643 (3197)	25251	577288
2	J. Topolsky	52	2057 (1980)	20858	264644
3	T. Ricker	42	4798 (4200)	36175	178198
4	N. Patel	42	3091 (2970)	26381	181447
5	P. Miller	40	5000 (4302)	30946	209522
6	D. Murph	34	11555 (10895)	63028	631108
7	C. Ziegler	33	1997 (1891)	14560	130673
8	R. Miller	27	1385 (1245)	9722	68795
9	V. Savov	26	944 (926)	8277	46436
10	D. Melanson	24	4856 (4196)	21878	147491

TABLE V
BLOGGERS RANKING ACCORDING TO VARIOUS METRICS

	Bloggers	BP	BI	$MEIBI$	$MEIBIX$
1	D. Murph	32	114	113	134
2	C. Ziegler	26	94	113	129
3	V. Savov	26	90	101	122
4	P. Miller	25	68	89	105
5	J. L. Flatley	24	58	55	67
6	T. Stevens	21	64	60	68
7	J. Stern	21	41	45	50
8	T. Ricker	20	79	92	107
9	J. Topolsky	20	72	113	125
10	R. Miller	20	55	77	85

crucial when studying blogosphere, since posts become outdated very quickly.

Now, let us apply our proposed methods on the acquired dataset in order to identify the bloggers who are both influential and productive. In our analysis, the performance of our methods is compared with the solution proposed in [11], the two time-sensitive metrics appeared in [12], and a ranking method which is a straightforward adaptation of a method coming from the bibliometric literature—the h-index [15]—(we call this method the plain method). According to this adaptation, the h-index for bloggers is defined as follows.

A blogger j has h-index h, if h of his/her N^j posts have received at least h citations each and the rest $N^j - h$ posts have received no more than h citations.

We divide the experimentation into two parts: in the first part, we compare the influential bloggers indicated by the proposed methods, to the bloggers found by the plain method: *MEIBI* and *MEIBIX*. We use the entire dataset as a baseline experiment to examine whether temporal considerations are worth examining. In the second part, we compare the influential bloggers indicated by the proposed methods with those found by the influence-flow method using the posts published in February 2010 to prove that even for small time intervals, the rankings are different.

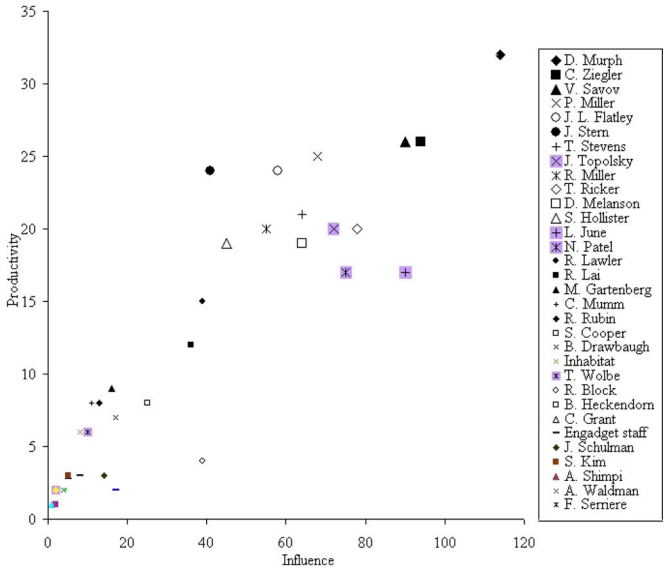

Fig. 1. Graphical representation of the behaviors of the Engadget bloggers.

A. *New Methods Versus the Plain One: MEIBI and MEIBIX*

Table IV presents a ranking of the ten most influential bloggers when the h-index [15] metric is used; recall that this metric examines the number of posts of each blogger and the number of incoming links to each posts, awarding both productivity and influence. The third column of Table IV displays the value of the h-index metric for each blogger whereas the next column shows the total number of posts he/she has submitted and how many of them have been cited by other posts, respectively. Finally, the last two columns illustrate the total number of incoming links and comments that all the posts of a blogger have attracted.

The information recorded in this table justifies that productivity and influence do not coincide. According to the h-index metric, the most influential blogger of Engadget is *R. Block* who has written 53 articles having at least 53 incoming links each. *R. Block* is the third most active blogger in Engadget; *D. Murph* and *P. Rojas* have submitted more posts. Although he has been inactive recently (his last post is dated eight months ago), he is still the most influential according to the h-index metric. This notification proves that h-index can indicate the most influential blogger, but cannot identify bloggers who are *both* influential and active *presently*.

In the following, we apply our proposed metrics to the posts of our dataset and we examine whether these metrics are suitable for the examined problem. Table V contains the ten top-ranked bloggers of Engadget along with the corresponding values of *BP*-index, *BI*-index, *MEIBI,* and *MEIBIX*.

The ranking provided by *BP*-index and *BI*-index does not agree with the rankings based on *MEIBI* and *MEIBIX* metrics (see Table V), and this is something that we partially anticipated: *MEIBI* does not account for *recent* influence, since it does not consider the age of the incoming links of each post. On the other hand, *MEIBIX* does not account for recent productivity, and furthermore, both metrics ignore the age of the comments made to each post, hence they do not award the recency of influence inside the communities. *J. Topolsky* of Engadget is one of the three top-ranked influential bloggers according to *MEIBI* and *MEIBIX* (values 113 and 125, respectively); however, our new methods consider him as the fifth most influential and the ninth most productive among the bloggers of the community. A similar disagreement between the rankings also holds for *T. Ricker*.

Fig. 1 visualizes the proposed metrics. Each point of the graph represents the blogging behavior of an individual blogger placed on the productivity–influence plane. The coordinates of each point are determined by the values of the *BP*-index and *BI*-index, and they reveal his/her productivity and influence. The points positioned near the top of the diagram represent bloggers with high productivity, whereas the points placed near the right boundary indicate highly influential

TABLE VI
BLOGGERS RANKING FOR FEBRUARY 2010 ACCORDING TO (FROM LEFT TO RIGHT) ENGADGET, INFLUENCE-FLOW MODEL, *MEIBI*, AND *MEIBIX*

	Bloggers	N	**Inlinks**	**Com**
1	D. Murph	172	1179	8203
2	V. Savov	146	1194	7518
3	D. Melanson	102	643	4716
4	C. Ziegler	101	1197	5839
5	T. Stevens	73	404	2699
6	R. Miller	70	454	2710
7	T. Ricker	58	641	3820
8	N. Patel	55	550	4350
9	P. Miller	52	445	2684
10	J. Flatley	41	196	1483

	Blogger
1	J. Topolsky
2	R. Lai
3	D. Murph
4	N. Patel
5	V. Savov
6	L. June
7	T. Stevens
8	D. Melanson
9	J. Stern
10	T. Ricker

	Blogger	m
1	C. Ziegler	45
2	D. Murph	42
3	V. Savov	41
4	D. Melanson	32
5	N. Patel	32
6	T. Ricker	32
7	P. Miller	26
8	R. Miller	24
9	T. Stevens	23
10	J. Stern	23

	Blogger	x
1	C. Ziegler	46
2	D. Murph	43
3	V. Savov	41
4	D. Melanson	33
5	N. Patel	32
6	T. Ricker	32
7	P. Miller	26
8	R. Miller	25
9	T. Stevens	23
10	J. Stern	21

bloggers. According to these notifications, a blogger is both influential and productive if the corresponding point is located near the top-right corner of the graph.

B. New Methods Versus the Influence-Flow Method

To compare the proposed metrics against the influence-flow method [11], we select a subset of the real data in order to conduct fairer experiments. Our intention is to discover whether these metrics can provide different rankings than those of the influence-flow method for a small period of time. For this reason, we selected to work upon the blog posts of February 2010 only. For comparison purposes, we also present in Table VI the top ten of active bloggers during February 2010 as this ranking is provided by the sites themselves.

In Table VI, we record the most influential and productive bloggers of Engadget for February 2010, as they are provided by the influence-flow method: *MEIBI* and *MEIBIX*. Neither *MEIBI* nor *MEIBIX* generates rankings that agree with the ranking provided by the blog sites themselves. For instance, Engadget concerns *D. Murph* as more influential than *C. Ziegler* in contrast to the rankings generated by these two metrics. Indeed, although *C. Ziegler* has authored fewer posts than *D. Murph*, his posts received more references from other posts. A similar notification can be made for *N. Patel*.

Now, let us examine how our proposed methods perform in this limited portion of the dataset. Fig. 2 provides an illustration similar to the one we provided in Fig. 1. The graph shows that six bloggers are both productive and influential, since the corresponding points are located near the top-right corner. The most influential blogger according to the *BI*-index is *C. Ziegler* in agreement to the rankings of *MEIBI* and *MEIBIX*. Although *D. Melanson* is not as productive as the other six, his posts gathered many incoming links and comments. In addition, there are five more bloggers that are highly productive, but their posts are not cited sufficiently.

The ranking generated by the influence-flow model considers *J. Topolsky* as the most influential blogger of Engadget during February 2010. This blogger has authored the best post of the entire community for February 2010, where he describes his impressions regarding Windows Phone 7 Series. The post has attracted 148 inlinks and 697 reader comments, whereas one outgoing link was included. The comparison of this ranking to the one produced by our methods indicates that *J. Topolsky* was not very productive during that period since he posted 17 times compared to the rich activity of *D. Murph*, who has published 172 posts. Regarding influence, we firmly believe that a single post is not a safe indication of a blogger's influence; the posts of other bloggers (i.e., *V. Savov* and *D. Melanson*) were cited and commented more frequently than those of *J. Topolsky*.

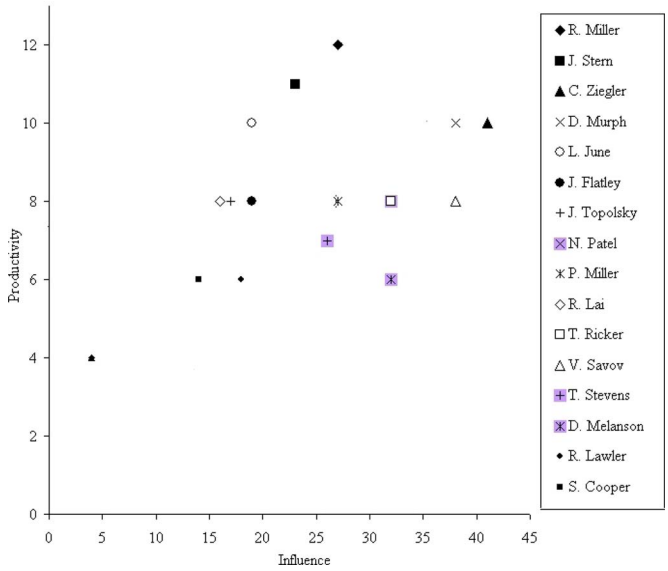

Fig. 2. Representation of the productivity and influence of the Engadget bloggers for February 2010.

IV. CONCLUSION

The blogosphere has recently become one of the most favored services on the Web. Its massive acceptance highlighted many interesting issues. This paper investigated the problem of identifying bloggers who are both influential and productive in a community blog site.

Based on the dynamic and time-varying characteristics of the blogoshpere, we introduced two metrics that attempt to address the problem in question. The main motivation for the introduction of these methods is that they can indicate individuals with potential commercial and advertising significance. The competing methods have not taken into account temporal aspects of the problem, which we argue are the most important ones when dealing with spaces like the blogosphere.

The first metric, i.e., *BP*-index, is used to evaluate the productivity of a blogger with respect to recency. The second metric, *BI*-index reflects the influence of a blogger inside and outside of a community by taking into consideration the number and the age of the incoming links and the comments. The combination of these two values is then used to characterize the bloggers. Hence, a blogger can be characterized as recently influential or recently productive, or both, or none.

These methods were evaluated against the state-of-the-art influential blogger identification methods, which have been reported in [11] and [12]. To evaluate our methods, we utilized data collected from Engadget, a real-world popular community blog site. The obtained results

verified that the new methods are able to identify significant temporal patterns in the blogging behavior and reveal some latent facts about the blogging activity.

REFERENCES

[1] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, "Why we blog," *Commun. ACM*, vol. 47, no. 12, pp. 41–46, 2004.

[2] [Online]. Available: http://technorati.com/blogging/feature/state-of-the-blogosphere-2009/.

[3] D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins, "The predictive power of online chatter," in *Proc. ACM Conf. Knowledge Discovery Data Min.*, 2005, pp. 78–87.

[4] N. Agarwal and H. Liu, "Blogosphere: Research issues, tools and applications," *ACM SIGKDD Explorations*, vol. 10, no. 1, pp. 18–31, 2008.

[5] K. E. Gill, "How can we measure the influence of the blogosphere?," in *Proc. Workshop Weblogging Ecosyst.: Aggregation, Anal. Dyn.*, 2004, pp. 1–5.

[6] A. Java, P. Kolari, T. Finin, and T. Oates, "Modeling the spread of influence on the blogosphere," in *Proc. ACM World Wide Web Conf.*, 2006, pp. 1–7.

[7] R. Cross, R. E. Rice, and A. Parker, "Information seeking in social context: Structural influences and receipt of information benefits," *IEEE Trans. Syst., Man, Cybern., Part C*, vol. 31, no. 4, pp. 438–448, Nov. 2001.

[8] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2005, pp. 207–214.

[9] Y. Zhou and J. Davis, "Community discovery and analysis in blogspace," in *Proc. ACM World Wide Web Conf.*, 2006, pp. 1017–1018.

[10] M. Louta and I. Varlamis, "Blog rating as an iterative collaborative process," *Semant. Adapt. Pers. Serv.*, vol. 279, pp. 187–203, 2010.

[11] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proc. ACM Conf. Web Search Data Min.*, 2008, pp. 207–218.

[12] L. Akritidis, D. Katsaros, and P. Bozanis, "Identifying influential bloggers: Time does matter," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, 2009, pp. 76–83.

[13] J. Zhan, C.-L. Hsieh, I.-C. Wang, T.-S. Hsu, C.-J. Liau, and D.-W. Wang, "Privacy-preserving collaborative recommender systems," *IEEE Trans. Syst., Man, Cybern., Part C*, vol. 40, no. 4, pp. 472–476, Jul. 2010.

[14] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proc. 3rd Text REtrieval Conf.*, 1994, pp. 109–126.

[15] Wikipedia. (Sep. 2010). The Hirsch h-index [Online]. Available: http://en.wikipedia.org/wiki/H-index.

[16] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos, "Generalized Hirsch h-index for disclosing latent facts in citation networks," *Scientometr.*, vol. 72, no. 2, pp. 253–280, 2007.