Investigating the impact of the blogosphere: Using PageRank to determine the distribution of attention

Lars Kirchhoff
Institute for Media and Communication Management
University of St. Gallen
St. Gallen, Switzerland
lars.kirchhoff@unisg.ch

Dr. Axel Bruns
Creative Industries Faculty
Queensland University of Technology
Brisbane, Australia
a.bruns@qut.edu.au
http://snurb.info/

Thomas Nicolai
Institute for Media and Communication Management
University of St. Gallen
St. Gallen, Switzerland
thomas.nicolai@unisg.ch

ABSTRACT

Much has been written in recent years about the blogosphere and its impact on political, educational and scientific debates. Lately the issue has received significant attention from the industry. As the blogosphere continues to grow, even doubling its size every six months, this paper investigates its apparent impact on the overall Web itself. We use the popular Google PageRank algorithm which employs a model of Web used to measure the distribution of user attention across sites in the blogosphere. The paper is based on an analysis of the PageRank distribution for 8.8 million blogs in 2005 and 2006.

This paper addresses the following key questions: How is PageRank distributed across the blogosphere? Does it indicate the existence of measurable, visible effects of blogs on the overall mediasphere? Can we compare the distribution of attention to blogs as characterised by the PageRank with the situation for other forms of Web content? Has there been a growth in the impact of the blogosphere on the Web over the two years analysed here? Finally, it will also be necessary to examine the limitations of a PageRank-centred approach.

INTRODUCTION

Much has been written in recent years about the blogosphere and its impact on political, educational, and scientific debates [1, 11, 14, 21]. Recently blogs and the blogosphere have received greater attention from the industry [29]. As the blogosphere still seems to be growing, even doubling its size every six month [37, 38], we conduct an analysis of its measurable impact onto the web itself. We use the popular PageRank algorithm that uses a model of Web use [7, 41] to measure the distribution of attention.

The paper is based on an analysis of the PageRank distribution for 8.8 million blogs in 2005 and 2006. PageRank is a proprietary algorithm used by Google in determining the relative importance of Websites; it is generated from Google's analysis of Websites and their patterns of linkage, and informs the results of Google searches by placing sites with a high PageRank closer to the top of results listings. While the exact details of the algorithm remain a trade secret, the rankings of individual Websites are readily available and were used in this study.

There have previous attempts by tools and services such as Technorati, BlogRunner, BlogLines, BlogStreets or BuzzMetrics to monitor the distribution of impact across the blogosphere. By contrast, this paper investigates blogs' impact not only on the limited realm of the blogosphere itself, but instead measures their influence on a large-scale basis, that is, on the Web in general: it focuses on the extent to which blogs have been able to gain a PageRank which indicates a positioning comparable to the sites of mainstream media or other key organisations. The presence of blogs with a high PageRank would point towards an influence well beyond what is occasionally described as the 'echo chamber' of the blogosphere itself.

This paper addresses the following key questions, therefore: How is PageRank distributed across the blogosphere? Does it indicate the existence of measurable, visible effects of blogs on the overall mediasphere? Can we compare the distribution of attention to blogs as characterised by the PageRank with the situation for other forms of web content? Has there been a growth in the impact of the blogosphere on the Web over the two years analysed here?

Finally, it will also be necessary to examine the limitations of a PageRank-centred approach. The PageRank itself can be seen as a measure of the average importance of a site in relation to the entire Web. It does not show the relevance of a given site for a specific topic determined by search keywords.

The remainder of this paper is organized as follows. In section II we begin with a review of research in blogs and the blogosphere as well as a literature review on current research in cybermetrics and graph theory applicable to our research questions to give an overview for further discussion. In section III we discuss our research approach; in particular the usage of Google's PageRank to analyze the impact of the blogosphere. In section IV we show how we gathered the data, and describe the results of our analysis. In section V we discuss the results, their implications, and their possible limitations. In the final section VI we draw the conclusions which can be derived from the results, and outline further directions for research.

II. BACKGROUND

Defining blogs and blogosphere: A representation of information and people

Blogs are typically building a corpus of regular, date-stamped entries which represent timeliness [15]. The traditional definition of this type of Webpage highlights their reverse chronological order and an archive functionality [5]. Blogs evolved out of the Web logs that were used as a way for individuals to comment on Web pages. They are mainly used for personally oriented written communication with the main aim to interact with readers in the distributed blogosphere [28]. In most cases blogs represent individual writers, whose texts exhibit both written and spoken qualities [31].

These specific linguistic properties as well as the temporal nature of blog entries and the structure of the blogosphere are the main differentiating factors in comparison to other types of Web content [28].

The blogosphere describes the entire network of blogs, and most research has been focused on the blogosphere and its properties. The structure of the blogosphere can be seen as a combination of a network of information and a network of people [22, 30]. These two types of networks can be identified through the different types of links that can be found in blogs. One type of link is commonly found in a list of links to the base URLs of blogs by friends of the blog author, or of other frequently read blog sites, known as a 'blogroll'; by equating the base URLs of blogs with their authors, such links thus construct a network of people. The second type of link is constituted by topical links in blog posts themselves, which are comparable to similar topical links found in other Websites. Some interaction links, such as

Trackbacks and links in comments, are very specific to blogs, and enhance the network of information as they attach more information to specific blog posts. In their operation, such links can also be seen as constructing a distributed conversation on specific topics, conducted across the blogosphere.

An important technology for the success of blogs and the dissemination of information within the blogosphere is RSS. This simple XML format makes it easy to keep track of changes in blogs and allows for easy content syndication and integration. This leads in some cases to fast responses and discussion within some parts of the blogosphere.

All of these properties are used as argument for the high impact of blogs mainly because these features help blogs to get indexed by Google very quickly. The very simple linking structure of blogs makes it easy for Google to retrieve the whole content from a blog. In combination with search engine-friendly URLs, almost all pages of blogs are indexed by Google [18]. This was a very unique feature at the beginning of the blog boom: many pages were indexed, and were linking to each other, creating the impression of a highly active, highly interlinked, highly influential network of sites. But does this really mean that a) interlinking within the blogosphere is stronger than it is for the rest of the Web, and that b) blogs have a higher impact on the overall mediasphere than other Web sources?

Related Work

As this paper examines the structural properties of a subset of the Web – the blogosphere – by using available data, literature in the field of cybermetrics, infometrics and Webometrics provides a valuable source.

Cybermetrics is derived from bibliometrical approaches and attempts to find methods and models to explain the intellectual structure of cyberspace. Early work in this field has been done by Larson's exploratory analysis of the intellectual structure of cyberspace [24], followed by Ingwersen [19], who proposed a new measure, the Web Impact Factor. This new measure enhances the Journal Impact Factor. Early results on the Web Impact Factor had been disappointing, due to the fact that search engines had been unreliable at the time. This situation has changed since then, as search engine data have become more stable for data collection [39].

Infometrics is an emerging area of quantitative studies of networks, which has gained more and more attention in recent years, exploring areas of automated Web issue analysis, citation analysis and word term analysis [12].

The methods used in cybermetrics are based on graph theory, which has been adapted to the specific properties on the Web lately. There is an extensive body of literature that examines the network graph properties of the Web for the purpose of finding useful measures and patterns [2, 8, 10, 20, 26].
Borrowing from these areas of interest, our work attempts to enhance knowledge in blogosphere research, which can be seen as part of both information science research and social sciences research, where the link structure and its implications have become the objects of study. There are several studies on social network properties of the blogosphere [3, 16, 22, 25]. Among these studies, Herring *et al*. [17] is widely recognized for its in-depth analysis of social network properties and its qualitative analysis of blog "conversations". Most of these studies, however, concentrate on analysis of the blogosphere itself and its structure, without considering implications for the Web as a whole. Most tools used to measure the blogosphere, including Technorati, BlogRunner, BlogLines, BlogStreets or BuzzMetrics [25], also track only the blogosphere.

Beyond such studies, some outstanding blog incidents (like "Dell hell" [36] or Kryptonite [35]) which are often cited and discussed in public media, and the still growing number of blogs, indicate a broader impact of the blogosphere on public consciousness. But how can we measure this, and how can we compare this impact to other types of online information sources?
We propose the use of a measurable figure that can be obtained for all information sources on the Internet, to make a comparison about the impact of certain types of Websites, such as blogs.

III. METHODOLOGY

Why use Google PageRank?

The PageRank function was presented in [7, 33] and is used by the commercial search engine Google. It is used to rank Web pages according to their PageRank values, which allows for the presentation of search results matching a query in decreasing order of their PageRank. PageRank values are stored as positive integer values, and Google exports PageRank values to the public to be used in the Google Browser plugin. Unfortunately, these values are only approximate values of the Google-internal PageRank calculations, both because they are exported to public

view only at set intervals and therefore do not show real-time ranking values, and because they are exported as logarithmic values ranging from 0 – 10.

Nevertheless, these values can be used to give estimation about the distribution of attention to any given Website. As the original intention of the PageRank function was to model the behavior of a random surfer browsing from page to page, the PageRank indicates the probability of a user visiting a certain page [4]. Additionally, PageRank values are global properties in terms of graph theory, as distinct from the in- or out-degree of a page, which is a local property [34]. Therefore, PageRank distribution is more reliable as a means of modelling the impact of the blogosphere on the whole Web than is mere degree distribution within the blogosphere.

This leads us to believe that PageRank value distribution can be used to model and compare the impact of the blogosphere as a whole on the rest of the Web. The distribution of PageRank within the blogosphere can be compared to the PageRank distribution of the whole Web [8, 10]. Although there are many other PageRank-style algorithms [2, 20], the Google PageRank is the most widely available page ranking scheme, and to our knowledge Google's index is the most comprehensive one currently available, thus allowing us to compare the blogosphere and the whole Web as widely as possible.

Nevertheless, there are a number of limitations to the use of PageRank. As mentioned above, one limitation is that Google only exports PageRank in logarithmic values. These values are only an approximate value of the actual PageRank used internally by the Google search engine. Furthermore, it is known that Google updates these export values only on a random, unknown basis, which means that the values gathered at any point may not reflect the correct, most recent PageRank at the time of gathering.

As PageRank is a computed figure, and more so as the available export values are only very coarse, it constitutes second-order data, with all its limitations. We do not know exactly how it is computed, even though the algorithm itself is partly known. As the granularity of available data is small, due to the very coarse export range of PageRank values, only rough estimates can be made.

We also assume that there are no modifications of the PageRank algorithm that are particular applied only to blogs - in other words, we assume that the PageRank of blogs is calculated in the same way as it is calculated for any other class of Websites.

Web Impact

Within this paper we define impact as the chance to be read and cited by many people, and therefore as the possibility to influence their cognitive reception and behavior. Impact reflects the ability of Websites and Webmasters to attract user attention [23]. At present, Google is known as the market leader for search engines in the world. According to Webhits, a large majority of people use Google when they search on the Web (in Germany, for example, about 85% of Internet users use Google as their primary search engine [42]). Therefore search engines act as gatekeepers for certain information. A site's position within results returned for any search is very important, as many people rely only on the results on the first pages [13]. Thus it can be concluded that PageRank is very important for the impact of any site in the Web. Pages with a higher PageRank tend to have higher visitor numbers, and it can be assumed that they therefore have more impact on ones attention.
This model of impact is related closely to similar discussions in the field measuring computer-mediated communication for scientific publications, which explore the usage of the PageRank algorithm for the Web pages of publications to identify the most influential publications [32].

IV. ANALYSIS

Data Gathering and Selection

According to Thelwall, there are different approaches to select an appropriate set of sites for surveys based on Web crawls [40].

- Using a copy of the whole database of a large search engine is the most complete approach, although it has its limitations due to the limitations of the search engine's completeness.
- Using the directory structure of a search engine or directory to select links
- Selecting pages from a walk within a predefined large initial set
- Using a meta crawler to combine the results from different search engines is another approach if the comparison of results is of most interest.
- Using a random choice of IP addresses
- Selecting Web sites by a systematic search of domain name space

We followed the second approach by systematically analysing a directory to retrieve a reasonable number of blog URLs. We chose blogger.com, a Google company, because it provides an easy to crawl URL structure and a considerable amount of listed profiles. Each profile is stored with a positive integer value such as http://www.blogger.com/profile/1. Each profile may contain one or more URLs to blogs that belong to the user, who owns the profile. We crawled over 15 million profiles (December 2005), from which we could extract 8,871,005 blog URLs. Many profiles could be found that have more than one blog URL. This indicates that a large number of empty profiles have been created without any actual blog URL, most probably only for testing purposes. As we intended to perform a comparative analysis, we chose the same profiles one year later to conduct to the same analysis. At this time (December 2006) we could extract slightly more blog URLs (8,888,523) from the profiles.

This selection is not complete by far, but includes most of the profiles and blog URLs available on blogger.com at the time of the first crawl. Technorati has announced in their current annual statistics that they now crawl more than 60 million blogs. Currently Technorati crawls about 92 million blogs, which means that our analysis covers slightly less than 10% of the whole blogosphere as it exists today. (At the time of our first crawl, the blogosphere was substantially smaller than it is now, of course.)

For each of these blog URLs, the Google PageRank was retrieved using the same API as is used by Google's browser plugin. Note that we only gathered the PageRank of the "root" or homepage of the blog, not the PageRank for all pages within the blog site.

Several implementations of the algorithms to retrieve the data are available. The latency time of fetching the results was the main issue here. Therefore we implemented a daemon script to run about 100 parallel processes, in order to keep the data gathering time very narrow and thereby avoid data artefacts from changes in the PageRank within a single crawl. The complete PageRank crawl of all blog URLs was conducted within a single day.

Results

Table 1 shows the results of both crawls. The first four columns show the number and percentage of blog URLs for a given PageRank value in 2005 and 2006. Column five shows the difference between 2005 and 2006 in total numbers, whereas column six [1] shows the subtractions between 2006 and 2005 percentage values. The last

column [2] describes the relative increase or decrease in 2006 compared to 2005 in percentage. This means nothing more than considering the number of PageRank occurrences from 2005 as the 100 percent baseline for 2006.

PR	# 2005	% 2005	# 2006	% 2006	# Diff 05/06	[1] % 05/06	[2] % 05/06
0	8.325.289	93,8483	8.469.444	95,2852	+144155	+1,4368	101,7315
1	129.825	1,4635	45.716	0,5143	-84109	-0,9491	35,2136
2	156.708	1,7665	101.159	1,1381	-55549	-0,6284	64,5525
3	136.454	1,5382	141.988	1,5974	+5534	+0,0592	104,0556
4	84.978	0,9579	92.078	1,0359	+7100	+0,0780	108,3551
5	29.695	0,3347	30.866	0,3473	+1171	+0,0125	103,9434
6	5.482	0,0618	5.595	0,0629	+113	+0,0011	102,0613
7	988	0,0111	580	0,0065	-408	-0,0046	58,7045
8	510	0,0057	392	0,0044	-118	-0,0013	76,8627
9	451	0,0051	330	0,0037	-121	-0,0014	73,1707
10	625	0,0070	375	0,0042	-250	-0,0028	60,0000

Table 1: blogosphere PageRank distributions 2005/2006

Figure 1 visualizes these results to get a better understanding how PageRank is actually distributed, and to identify anomalies more easily. As the results show a large value range, we used a logarithmic scale to maintain readability. Figure 1a displays the distribution of our first crawl in 2005 and Figure 1b the one we conducted in 2006.

a) b)

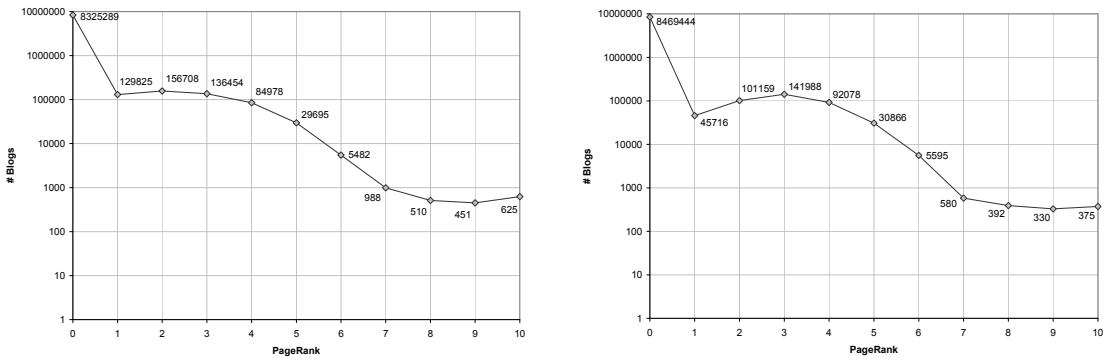

Figure 1: Blogosphere PageRank distribution 2005 & 2006

As both graphs above look almost identical with only minor changes, we needed another graph to examine whether these changes are significant for any particular

PageRank. Figure 2 shows the change in percentage for each PageRank, as a percentage of the all 8.8 million blogs covered by our study, and highlights that from this point of view, the most drastic changes happened for lower PageRanks – in particular for PageRanks between 0-2. This seems natural as the total number of blogs involved is higher and therefore more changes happen.

However, this graph only shows part of the full story. Figure 3 indicates that developments at each individual PageRank level are vastly different: the number of blogs at PageRank 0 remained relatively steady between 2005 and 2006, and blog numbers at PageRanks 3-6 have grown slightly from one year to the other, but the number of blogs at PageRanks 1 and 2, and at PageRanks 7 to 10, has diminished substantially; indeed, PageRank 1 now contains only just over a third of the number of blogs which existed at that level one year earlier. We discuss possible explanations for these substantial, inconsistent changes below.

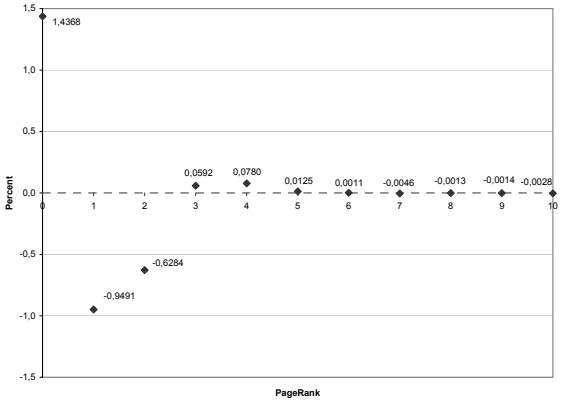

Figure 2: Change of PageRank distribution between 2005 and 2006 (% of all blogs)

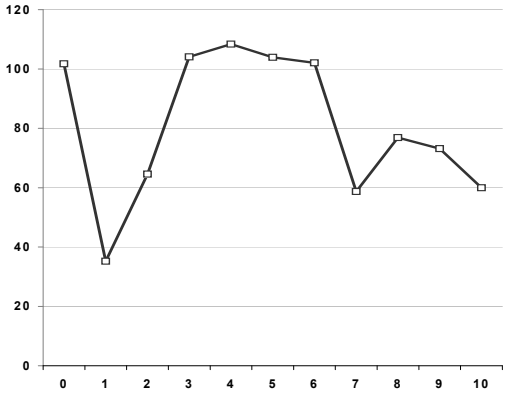

Figure 3: Change of PageRank distribution between 2005 and 2006 (% of blogs at the same PageRank level)

V. DISCUSSION

The distribution of PageRank of the blogs follows the skewed distribution of scale-free networks as described by Barabasi [6]. As PageRank is based on a user model which includes graph analysis of the incoming and outgoing links, such a result was to be expected. One of the characteristics of scale-free networks is that the distribution of outgoing and incoming links follows a power law [6]. As seen in figure 1, the PageRank distribution of the blogosphere is of similar character.

There are some anomalies from a power law distribution at a PageRank of one and two – both in 2005 and 2006. It is possible that such anomalies are caused in part by Blogger's placement as a market leader in blog hosting: as a first port of call for many users dabbling in blogging, it may experience unusually high levels of user churn at the very bottom end of the PageRank scale, which appear in our graphs as anomalous power law patterns.

Furthermore it can be observed that the graphs (figures 1a and 1b) show rather high values for a power-law distribution at a PageRank of eight to ten. This might be caused by a higher number of links within the blogosphere, where hubs as described by Barabasi [6] are even more significant than in other networks. In other words, hubs are linked comparatively more within the blogosphere than they are in the rest of the Web; this is observable for example in the common use of the "via" link, which is used to reference where a piece of news has been found. Many bloggers commonly engage in a process of coping news items from key blog hubs and adding their own comments to them; in most cases this is done to let friends within the local peer network know what is interesting in the wider Web, while giving credit to the source.

Apart from this, the impact of blogs within the blogosphere (measured in terms of PageRank) seems to be largely similar to that of other classes of Websites on the Internet. There is only a small number of high-influence blogs which command high levels of attention, complemented by a much larger group of smaller blogs which do not have any real impact beyond their local peer networks. The frequently emphasized effect of blogs as building strong link relationships appears to be less pronounced than it is assumed to be in many publications: while a significant degree of interlinkage takes place within the blogosphere network through the operations of posting, commenting, trackbacks, and blogroll linking, this does not appear to significantly increase the impact of the blogosphere as a whole, as our findings suggests.

What the changes across the PageRank levels between 2005 and 2006 show is a marked contraction of attention at the higher PageRank levels to a smaller number of blogs – in other words, the formation of a smaller and more entrenched 'A-list' of bloggers in our sample. If such tendencies are symptomatic for the wider blogosphere beyond our sample, they would indicate a growing focus of the blogosphere on a handful of key opinion leaders, even in spite of the continued

increase in the overall number of blogs now available. Furthermore this evidence provides some similarities with the Matthews effect [9, 27] in science, who describes an asymmetric distribution of attention across scientific authoring.

At the same time, there is also some pronounced growth in the middle of the PageRank continuum (levels 3 to 6), driven perhaps by the majority of consistent bloggers who attract a solid number of users and develop a strong local network of peers, but do not manage to rise beyond this to the national and global public attention indicated by higher PageRank ratings. Another, systemic reason for the decline at levels one and two and the increase at levels three to six may also be the fact that blogs are relatively quickly included in blog directories and other aggregator Websites (such as Technorati); perhaps more quickly than other classes of Websites are highlighted in their fields. It may therefore be the case that through such automatic linking by key sites, blogs gain a higher PageRank in this lower segment of the PageRank scale relatively soon, in comparison to other Websites.

The simultaneous decline in the number of blogs with a PageRank of one or two which we have observed is not compensated through the growth in blogs with a PageRank with three to six, however – this could also point to an attrition of bloggers, or at least to the attrition of bloggers *away from Blogger* (and thus out of our sample) and to other, less generic and mass market-oriented blog hosting services, as their practice develop and mature. If such assumptions of broader diachronic tendencies are correct (and further in-depth analysis of our data will be required to identify them reliably), then this could also explain the small but continuing growth at PageRank level zero even in spite of the sharp decline at levels one and two: here, we may see the emergence of a new generation of would-be bloggers, who begin their blogging by dabbling in Blogger, but possibly move on from there rather than stick with the site as their blogging practice matures.

The advantage of the use of PageRank as calculated by Google is that it takes into account usage patterns on the Web as a whole, and not only the distribution of links within the blogosphere or a custom sample of Web pages. This provides a better, more global estimation of the distribution of impact than do observations relying only on link distribution within the blogosphere.

VI. CONCLUSION

Google PageRank is a widely available figure, which can be easily obtained. As it is used to order the result of the most frequently used search engine, it can be assumed that the PageRank of a Web page has significant influence on the attention a Web page perceives from the overall population of users. In spite of the limitations discussed above, it can be used to generally measure the influence and impact of specific segments of the Web. Furthermore, it can be used as an indicator for structural properties of these segments.

Google's PageRank is not the only available source of data on the importance of Web pages, of course. Alexa, an Amazon company which collects Web usage data via a browser plugin, is another provider of data which may be interesting for our purposes. As its plugin tracks the pages a user visits, Alexa is able to gather usage statistics for any given page on the Web. On that basis, Alexa is calculating an overall Alexa Rank: the most popular page on the Web is ranked at an Alexa Rank of one. It would be interesting to compare the results of our PageRank analysis with the distribution of the Alexa Rank.

Another comparable figure is the Technorati Rank, which focusses specifically on ranking pages in the blogosphere for their 'authority' (measured in terms of in- and outlinks). Technorati is the key search engine for the blogosphere, indexing and ranking more than 92 million blogs at present. It would be interesting to collect the Technorati Rank for blogs with a PageRank higher than six, and to compare these two figures.

Beyond the establishment of PageRank patterns for the blogs hosted by Blogger (which we hope constitute a representative sample for the wider blogosphere), further research will be needed to examine the changes from year to year. As noted, it will be interesting to examine whether the top end of the PageRank scale is populated by a relatively steady set of 'A-list' blogs (and whether the decline in numbers at such levels is therefore a sign of a further contraction of attention – a further vertical differentiation even within the 'A-list'), or whether there are significant blog churn and attention shifts even at this level; similarly, it would be useful to trace whether and how (and perhaps most importantly, what percentage of) different cohorts of bloggers gradually advance from lower levels in the

PageRank continuum to positions of greater visibility, influence, and impact. Conducted over time, this would help identify the processes of large-scale communal evaluation and filtering which lead to the gradual emergence and rise of influential new bloggers; it may also lead to the identification of distinct generations of users publishing their work through Blogger and other services.

REFERENCES

[1] A. Adamic, L. and N. Glance. *The political blogosphere and the 2004 U.S. election: divided they blog*. in *Conference on Knowledge Discovery in Data, Proceedings of the 3rd international workshop on Link discovery*. 2005. Chicago, Illinois, USA: ACM Press

[2] Abiteboul, S., M. Preda, and G. Cobena. *Adaptive On-Line Page Importance Computation*. in *International World Wide Web Conference*. 2003. Budapest, Hungary.

[3] Bachnik, W., et al., *Quantitive and sociological analysis of blog networks.* Acta Physica Polonica, Series B, 2005. **36**(10): p. 3179-3191.

[4] Baeza-Yates, R., C. Castillo, and V. López, *Characteristics of the Web of Spain.* International Journal of Scientometrics, Informetrics and Bibliometrics, 2005. **9**(1).

[5] Baoill, A.Ó., *Conceptualizing The Weblog: Understanding What It Is In Order To Imagine What It Can Be.* Interfacings: A Journal of Contemporary Media Studies, 2005.

[6] Barabási, A.-L., R. Albert, and H. Jeong, *Scale-free characteristics of random networks: the topology of the world-wide web.* Physica A, 1999. **281**: p. 69-77.

[7] Brin, S. and L. Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. in *7th WWW conference*. 1998. Brisbane, Australia: Elsevier Science.

[8] Broder, A., et al. *Graph structure in the web*. in *9th International World Wide Web Conference*. 2000. Amsterdam.

[9] Cole, S., *Professional Standing and the Reception of Scientific Discoveries.* American Journal of Sociology, 1970. **76**(2): p. 286-306.

[10] Donato, D., et al., *Large scale properties of the Webgraph.* The European Physical Journal B - Condensed Matter, 2004. **38**(2): p. 239-243.

[11] Drezner, D. and H. Farrell, *The power and politics of blogs*, in *American Political Science Association*. 2004.

[12] Egghe, L., *Editorial: Expansion of the field of informetrics - The second special issue.* Information Processing and Management 2006. **42**: p. 1405–1407.

[13] Fallows, D., *Search Engine Users*. 2005, Pew Internet & American Life Project: Washington. p. 36.

[14] Farmer, J. and A. Bartlett-Bragg. *Blogs @ anywhere: High fidelity online communication*. in *ASCILITE*. 2005. Brisbane.

[15] Gillmor, K.E. *How can we measure the influence of the blogosphere?* 2004 [cited 2006 14.01.2006]; Available from: http://faculty.washington.edu/kegill/pub/www2004_keg_ppt.pdf.

[16] Gruhl, D., et al. *Information Diffusion Through Blogspace*. in *WWW Conferences Archive*. 2004.

[17] Herring, S.C., et al. *Conversations in the Blogosphere: An Analysis "From the Bottom Up"*. in *38. Hawaii International Conference on System Sciences (HICSS-38)*. 2005. Hawaii: IEEE Press.

[18] Hiler, J. *Google loves Blogs - How Weblogs Influence A Billion Google Searches A Week.* 2002 [cited 2006 16.01.2006]; Available from: http://www.microcontentnews.com/articles/googleblogs.htm.

[19] Ingwersen, P., *The calculation of Web Impact Factors.* Journal of Documentation, 1998. **54**(2): p. 236-243.

[20] Kleinberg, J., *Authoritative sources in a hyperlinked environment.* Journal of the ACM, 1999. **46**(5): p. 604-632.

[21] Kolari, P., A. Java, and T. Finin. *Characterizing the Splogosphere*. in *3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th International World Wid Web Conference*. 2006. Edinburgh, UK: University of Maryland, Baltimore County.

[22] Kumar, R., et al. *On the bursty evolution of Blogspace*. in *6th International WWW Conference*. 2003. Budapest, Hungary.

[23] LaBerge, D., *Attention, Awareness, and the Triangular Circuit.* Consciousness and Cognition 1997. **6**(2-3): p. 149-181.

[24] Larson, R.R. *Bibliometrics of the World Wide Web: An Exploratory Analysis of the Structure of Cyberspace*. in *ASIS '96 Proceedings of the 59th ASIS Annual Meeting*. 1996. Baltimore, USA: Information Today.

[25] Marlow, C.A. *Linking without thinking: Weblogs, readership, and online social capital formation*. in *International Communication Association Conference*. 2006. Dresden.

[26] McIlraith, S.A. and D.L. Martin, *Bringing semantics to Web services.* IEEE Intelligent Systems, 2003. **18**(1): p. 90-93.

[27] Merton, R.K., *The Matthew Effect in Science.* Science, 1968. **159**(3810): p. 56-63.

[28] Mishne, G., *Information Access Challenges in the Blogspace*, in *IIIA-2006: International Workshop on Intelligent Information Access*. 2006: Helsinki, Finland.

[29] Nagel, S.C., *Using the world of blogs for project and financial management.* The Bottom Line: Managing Library Finances, 2004. **17**(3): p. 102-105.

[30] Newman, M.E.J., *The Structure and Function of Complex Networks.* SIAM Review, 2003. **45**(2): p. 167-256.

[31] Nilsson, S., *The Function of Language to Facilitate and Maintain Social Networks in Research Weblogs*. 2003, Umea Universitet.

[32] Noruzi, A., *The Web Impact Factor: a critical review.* The Electronic Library, 2006. **24**(4): p. 490 - 500.

[33] Page, L., et al., *The PageRank Citation Ranking: Bringing Order to the Web*. 1998, Computer Science Department: Stanford.

[34] Pandurangan, G., P. Raghavan, and E. Upfal. *Using PageRank to Characterize Web Structure*. in *8th Annual International Computing and Combinatorics Conference (COCOON)*. 2002.

[35] Pikas, C.K., *BLOG searching for competitive intelligence, brand image, and reputation management.* Online (Wilton, Connecticut), 2005. **29**(4): p. 16-21.

[36] Sentinel, M., Onalytica, and I. Future (2005) *Measuring the influence of bloggers on corporate reputation*.

[37] Sifry, D. *State of the Blogosphere, April 2006 Part 1: On Blogosphere Growth*. [Blog Post] 2006 [cited 2006 12.11.2006]; Available from: http://www.sifry.com/alerts/archives/000432.html.

[38] Sifry, D. *The State of the Live Web, April 2007*. [Blog Post Website] 2007 [cited 2007 06.04.2007]; Available from: http://www.sifry.com/alerts/archives/000493.html.

[39] Thelwall, M., *A comparison of sources of Links for academic Web Impact Factor Calculations.* Journal of Documentation, 2002. **58**(1): p. 60-72.

[40] Thelwall, M., *Methodologies for crawler based Web surveys.* 2002.

[41] Thelwall, M., *Can Google's PageRank be used to find the most important academic Web pages?* Journal of Documentation, 2003. **59**(2): p. 205-217.

[42] Webhits. *Web Barometer*. 2006 [cited 2006 07.06.2006]; Available from: http://www.webhits.de/deutsch/index.shtml?/deutsch/webstats.html.