Who Blogs What: Understanding Behavior, Impact and Types of Bloggers

Kang Zhao[1], Akhil Kumar[2], Massimiliano Spaziani[3], John Yen[1]

[1]College of Information Sciences & Technology, Penn State University, University Park, PA 16802, USA
[2]Smeal College of Business, Penn State University, University Park, PA 16802, USA
[3]Department of Internet Media & Digital Communication, Telecom Italia, Rome, Italy

{kangzhao, akhilkumar}@psu.edu, massimiliano.spazianibrunella@telecomitalia.it, jyen@ist.psu.edu

Abstract

We investigated bloggers' publishing patterns by focusing on the topics that their posts cover. Applying clustering algorithms on the dataset from a blog website of 370,000 posts from 2,275 blogs, we identified two types of bloggers: *specialists* and *generalists*. Then we compared their respective contributions to the blogosphere in terms of productivity and buzz-factor. Our analysis suggests that specialists generally have a higher impact than generalists. It also reveals that among specialists, only a small fraction create a large "buzz" or produce a voluminous output.

1. Introduction

Blogs (weblogs) have become an important online media to publish, share, and disseminate information on the Web. Some blogs (e.g. Politico.com) have had a major impact on government and corporate policy, and have become a must reading for officials. In the business world, the analysis of blogs can help to identify what bloggers may want to purchase [1]. Corporations can also leverage blogs to market their products and to interact with existing and potential customers [2, 3]. *Blogosphere*, the world or community of blogs, is growing rapidly and has attracted a lot of researchers from different disciplines. Previous research on blogosphere has studied the topology and evolution of blogosphere [4], citation networks among blogs and posts [5], information propagation through blogs [6], reading behaviors of blog users [7], etc.

In this study, we focused on the publishing patterns of bloggers of both special- and general-interest blogs. While some research classifies political bloggers at the micro level (e.g. liberals versus conservatives) based on the political issues covered in their posts [8], there is little research that studies bloggers' publishing patterns at a broader topic level, such as sports, technology, etc. Here, we study the publishing behavior and impact of bloggers by macro-level topics and also some general topics. We believe that revealing which topic a blogger would like to cover and comparing impacts of bloggers who have different topical interests will improve our understanding of bloggers' behavior and contributions. Such understanding may inform the design of the blogosphere for bloggers and readers, improve the effectiveness of online advertising for advertisers, as well as help the utilization of blogs in other areas.

We briefly introduce our dataset and perform a preliminary analysis in Section 2. Then, Section 3 describes how we cluster bloggers using the topical distribution of their posts and compare contributions of different types of bloggers. The differences among the contributions from bloggers of the same type are studied. The paper concludes with a discussion of future work.

2. Dataset and preliminary analysis

Our research is based on data from the Italian blog website BlogNation (www.blognation.it), which covers a broad range of topics. The dataset contains more than 370,000 posts, published between December 2009 and May 2010, from 2,275 blogs. Although bloggers did not specify the topic category of their blogs, BlogNation extracted the content of their posts and used the natural language analysis tool Cogito [9] to classify the topic of each post into 8 topic categories: *news, cars, culture, entertainment, food, sports, technology,* and *others*.

Many parameters of the dataset follow a Power Law ($P(k)=k^{-r}$ [10]), or a similar distribution, with long right-tails. For example, the number of comments a post receives follows a near power-law distribution (see Figure 1) with $r \approx 2.1$, meaning that most receive none or a few comments, while a select few draw many comments. The number of posts of a blogger (most publish more than 10) also accords with a near Power Law distribution (see Figure 2) with $r \approx 1.3$.

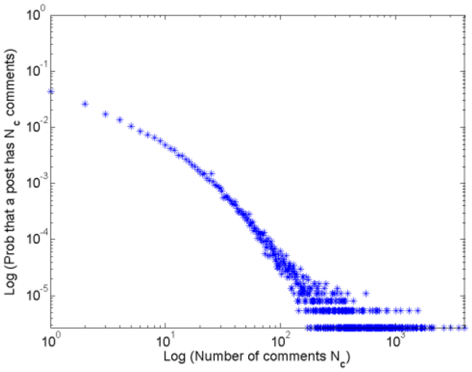

Figure 1. The log-log distribution of the number of comments for a post.

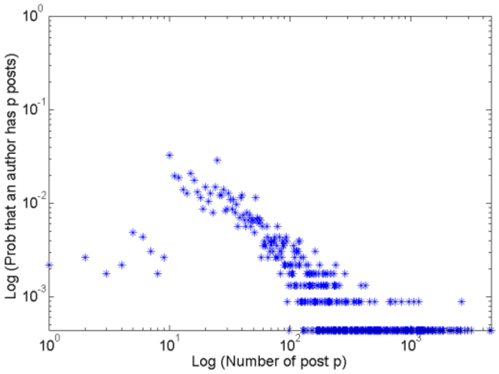

Figure 2. The log-log distribution of the number of posts a blogger published.

We constructed a post-post citation network, where posts represent nodes and edges denote the citation relationship among posts. The in-degree of a node is the number of other posts that cite this post, while the out-degree is the number of times this post cites other posts. This network is sparse because most posts do not cite others or get cited. 43,047 posts from 906 blogs have non-zero in- or out-degrees, and are connected by 50,434 edges in the network. The distributions of in- and out-degree generally follow Power Laws. However, there is no giant component (a connected sub-network that contains a majority of all the nodes). Instead, the 43,047 nodes are divided into 9,754 components, which are disconnected from each other. The number of nodes in (or the size of) each component also follows a near power-law distribution.

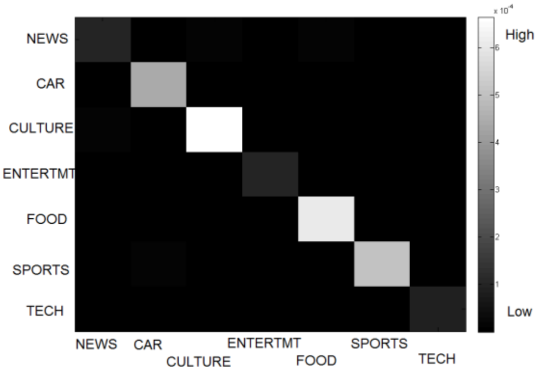

Figure 3. Topical citation densities in the post-post citation network.

Table 1. The DBIs for clustering with different k values.

k	**DBI**		**k**	**DBI**
2	1.2104		11	0.6309
3	1.1197		12	0.7380
4	1.1107		13	0.7564
5	1.0462		14	0.7527
6	0.8919		15	0.8038
7	0.6378		16	0.8448
8	0.6392		17	0.8786
9	0.5269		18	0.8966
10	0.6134		19	0.9045

It was found that citation among posts exhibits topically assortative patterns [11] as a post tends to cite another post within the same topic. We represent the *topical assortativity* with the cross-topic citation density. The *citation density* between a pair of topics (X,Y) measures how likely a citation link exists between a post on topic X and another on topic Y, and is defined as:

$$D_{XY}=2\times C_{XY}/P_X\,P_Y\,,$$

where C_{XY} is the number of citation links between a post on topic X and another one on topic Y; P_X and P_Y are the total number of posts on topic X and Y, respectively.

We illustrate the topical citation densities with a density map in Figure 3, each cell representing the citation density between two topics. As the “others” category does not really correspond to a specific topic, we only include 7 topic categories in the map. The density values along the diagonal range from 8.7×10^{-5} to 66×10^{-5}, and are much higher than those in the cells off the diagonal. For instance, culture posts are 30 times more likely to cite other posts within culture than posts on other topics; however, this is less so for news posts, where same-topic citation is only 3 times more likely than cross-topic citation. We conjecture that this is because the coverage of news posts is generally broader than of other topics, and so they have a less incestuous tendency. A rather narcissistic trend observed in the dataset relates to self-citation. Among all the 50,434 citation links, 86% of the time the bloggers cite their own blog posts!

Next we turn to a more in-depth analysis of the BlogNation dataset.

3. Clusters, impact, generalists and specialists

3.1. Clustering of bloggers

We first represented the topic profile of each blogger *i* with a *topic vector* $T_i=<t_{i1}, t_{i2}, ..., t_{i8}>$, where t_{ij} ($j=1,2,...8$) represents the percentage of blogger *i*’s posts on topic *j*. For example, if a blogger has published 10 posts, 2 on news and 8 on technology, her topic vector will be *<0.2, 0, 0, 0, 0, 0, 0.8, 0>*.

Then the bloggers were clustered on the basis of their topic vectors using the well-known *k-means algorithm*. This algorithm partitions all the bloggers into *k* clusters so that “similar” bloggers belong to the same cluster. While alternative clustering algorithms exist, the simple k-means worked well on our dataset and generated reasonable clusters of bloggers. To find the best *k* for our dataset, we tried different *k* values (from 2 to 20) and then used the Davies-Bouldin Index (DBI) [12] to evaluate the quality of clustering results. Briefly speaking, DBI is based on a compactness measure of clusters divided by an inter-cluster distance measure. On one hand, DBI favors smaller clusters because the intra-cluster distance is lower in a smaller cluster. On the other hand, it also penalizes short inter-cluster distances so that partitioning the data into a large number of small clusters that are very close to each other is also discouraged. The solution with the lowest DBI gives a balanced clustering. For our dataset, the DBI suggests $k=9$ (see Table 1).

Among the 9 clusters found, 7 are topic-specific clusters, as there is a one-to-one mapping between the 7 topics and the corresponding clusters. Bloggers in a *topic-specific cluster* are found to publish more than 90% of their posts on one topic alone. For example, one cluster of 278 bloggers focuses heavily on technology, because, on average, 95.4% of their posts are about technology. For sports bloggers in a 147-blogger cluster, the average percentage of sports posts is 98%. Similarly, we also find clusters for entertainment, food, news, cars and culture. Because the 1486 bloggers (about 65% of all bloggers in BlogNation) in the 7 topic-specific clusters publish posts mainly on a single topic, we call them *specialists*.

In contrast to the 7 topic-specific clusters, the other two clusters do not exhibit such a strong focus on one specific topic. For example, a cluster of 423 bloggers published 36% of their posts in news, 11% on entertainment and 28% on other topics. This means bloggers in the two clusters cover a broader range of topics in their posts than specialists do. Thus, we combine the two clusters and classify the 789 bloggers (about 35% of all bloggers) in the two clusters as *generalists*.

3.2. Impact metrics

Before investigating how specialists and generalists contribute to the blogosphere, metrics are needed to measure a blogger’s impact. Many factors could reflect the impact of a blogger, but no single factor can solely represent such impact. Thus a blogger’s impact in the blogosphere is often approximated by combining multiple factors, such as the number of posts, the length of posts, the number of citations, etc. [13]. While some bloggers may abuse the metric of impact to boost their impact ranking by publishing spam posts and comments, the consideration of this type of behavior is beyond the scope of this research. On the basis of data availability, and for simplicity, we chose two straightforward metrics: productivity and buzz-factor.

Productivity is simply the total number of posts a blogger publishes in a given period. This is a good measure of quantity but does not reflect quality. Thus we introduce another metric called buzz-factor. *Buzz-factor* is a measure of the buzz a post generates. We approximate this metric with the number of comments for a post. As suggested in previous research [13], a post that can attract readership and generate discussion among readers will likely receive many comments. As mentioned in Section 2, the number of comments received for posts by a blogger follows a highly-skewed power-law distribution. Therefore, we consider only the top-N most commented posts (MCPs) of a blogger (e.g. “Top-1”, “Top-5”, “Top-10”), and average across them to determine the blogger’s buzz-factor.

Generalists vs. specialists: Figure 4 compares the distribution of the total number of posts for generalists and specialists. The approximate Power Law curve for specialists (*r≈1*) lies above the one for generalists (*r≈1.2*), and its slower decay suggests that specialists are generally more productive than generalists. Figure 5 compares the average number of comments for a blogger’s *top 5 most cited posts*. The two Power Law curves in this figure show that specialists (*r≈0.9*) tend to attract more comments than generalists (*r≈1.3*), thus outperforming generalists on both metrics, and indicating their larger impact.

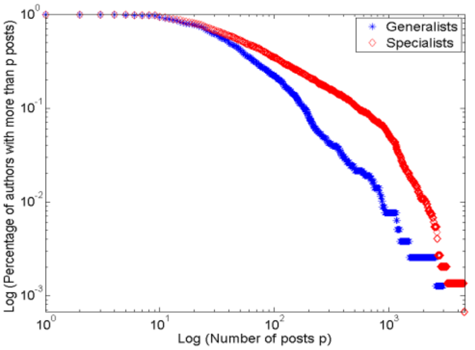

Figure 4. Cumulative distributions of the total number of published posts

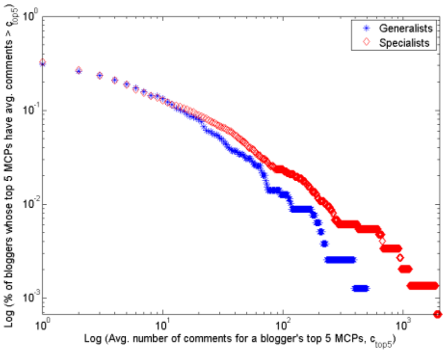

Figure 5. Cumulative distributions of the average number of comments for the top 5 MCPs.

One might hypothesize that the difference in contribution reflects the difference between professional and amateur bloggers. Compared with generalists, specialists tend to be more professional bloggers who publish more regularly and thus contribute more posts. They also bring more expertise and dedication to their topic, making their posts more useful. Generalist blogs, on the other hand, tend to come from more amateur bloggers who publish posts on more than one topic of general interest to them. Hence, their posts may tend to lack depth.

3.3. A drill-down into specialist bloggers

Our dataset has more specialists than generalists, and, as noted above, specialists tend to have a greater impact than generalists when measured with our impact metrics of productivity and buzz-factor. Along the lines of reference [13], which identified some “inactive but influential” and “active but non-influential” bloggers, a next logical step is to understand if there are different types of specialists and how their contributions differ from each other. As an example of a specialists group, we decided to focus on the 441 “news” bloggers, who constitute the largest topic-specific cluster in our topic-based clustering of bloggers.

Our analysis suggests that various “news” bloggers do contribute in different ways. The scatter plot of the two metrics in Figure 6 and a correlation analysis reveal that there is no strong correlation between the two metrics. A highly productive news blogger is not necessarily one with a high buzz-factor, and vice-versa. In addition, while the productivity of news bloggers has a larger variance, their buzz-factor is more closely scattered with fewer outliers.

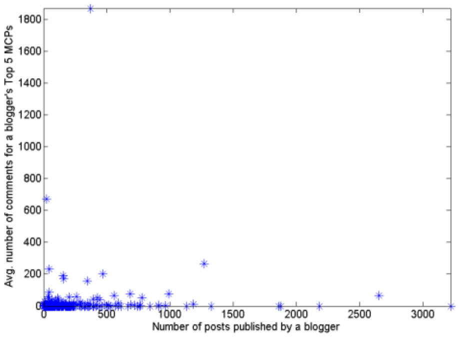

Figure 6. The scatter plot of news bloggers: total number of published posts vs the avg. number of comments for a blogger’s top 5 MCPs.

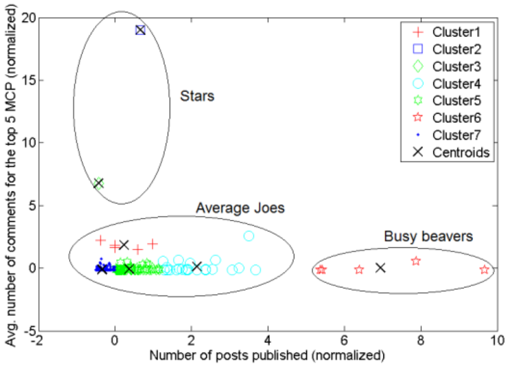

Figure 7. Clustering results with k-means ($k=7$). Axes denote the normalized values. Circles contain clusters aggregated into the same subgroup.

To determine subgroups among news bloggers, we again used k-means clustering. However, instead of clustering by a blogger’s topic distribution, it was performed by the bloggers’ productivity and buzz-factor. In other words, each blogger i is represented by tuple $B_i=<P_i, Z_i>$, where P_i represents the blogger’s productivity, i.e., the total number of published posts; and Z_i the blogger’s buzz-factor, measured by the average number of comments on the blogger’s top 5 MCPs. As P_i and Z_i have different scales, which may bias the calculation of Euclidean distance in the k-means algorithm, we normalized them before performing the clustering. Figure 7 illustrates the clustering results for $k=7$. As both axes are normalized, the values on each axis denote the number of standard deviations away from the mean.

To make the partitioning intuitively easier to understand, we condensed the 7 clusters into three subgroups (see circles in Figure 7) and labeled them suitably. Table 2 shows the subgroups and the number and percentage of bloggers in each one. As one might expect, *most bloggers have only an average productivity and buzz-factor*. There are considerably fewer bloggers with a high buzz-factor or productivity. We also observe that no blogger excels on both the metrics, illustrating the common quantity-quality tradeoff.

Table 2. The number of news bloggers in each subgroup.

	Average buzz	High buzz
Average productivity	434 (*Average Joe*, 98.4%)	2 (*Star*, 0.5%)
High productivity	5 (*Busy beaver*, 1.1%)	0

4. Conclusions

By studying the publishing patterns of bloggers and their impact in the blogosphere, we gained several insights. Using a large dataset from an Italian blog website, we were able to first partition bloggers into clusters, and distinguish between specialist and generalist bloggers based on the topic vectors of their posts. Their impact in the blogosphere in terms of productivity and buzz-factor, suggests that specialists made more useful contributions than generalists, perhaps from their deeper subject matter expertise. Further analysis of a group of specialists revealed that their contribution styles are different: some are very productive but generate average buzz ("busy beavers"), while others create a lot of buzz from only a few posts ("stars"). The results of a previous study on the activity and influence of bloggers in a technology blog website [13] were similar, but we did not find a category that combines busy beavers and stars.

This research helps to understand the behavior and contributions of bloggers in the blogosphere. Specifically, it can help a website such as BlogNation to decide which bloggers contribute more, and whose posts to display on the front page of the website. The findings may also be used by blog search engines to tag and rank bloggers. There are implications as well for improving click-through rates in online advertising. For example, a sponsored search service such as Google AdWords may assign higher scores to blogs with high buzz-factors because these blogs have a track record of attracting more eyeballs. Furthermore, a blog service provider is able to deploy advertisements in a more targeted fashion, e.g., place automobile ad banners in the blogs of a car specialist. Finally, the study may also provide insights for finding "important" bloggers, with possible implications in marketing, public relations, political campaigns, etc.

There are still many unanswered questions for the future. We would like to explore the temporal publishing patterns of bloggers, such as their times of publishing posts and temporal intervals between their posts. We conjecture that specialists and generalists may have different temporal publishing patterns. For example, specialists may publish more regularly and frequently on weekdays, and less on weekends and holidays. Generalists might have a more sporadic pattern. Also, if more data becomes available, we plan to study the impact of bloggers using networks based on various relationships such as citation, comment, trackbacks and blogrolls [7]. This will allow us to devise a more comprehensive buzz-factor metric. We also plan to analyze other blogospheres, and with other clustering techniques, and thus generalize our findings further.

References

[1] G. Mishne and M. d. Rijke, "Deriving wishlists from blogs: show us your blog, and we'll tell you what books to buy," in *the 15th Intl. Conference on WWW*, Edinburgh, Scotland, 2006, pp. 925-926.

[2] J. Wright, *Blog Marketing: The Revolutionary New Way to Increase Sales, Build Your Brand, and Get Exceptional Results*: McGraw-Hil, 2005.

[3] R. Scoble and S. Israel, *Naked conversations: how blogs are changing the way businesses talk with customers*: John Wiley and Sons Ltd, 2006.

[4] R. Kumar, *et al.*, "Structure and evolution of blogspace," *Comm. of ACM*, vol. 47, pp. 35-39, 2004.

[5] J. Leskovec, *et al.*, "Patterns of Cascading Behavior in Large Blog Graphs," in *SIAM International Conference on Data Mining*, Minneapolis, MN, 2007, pp. 551-556.

[6] D. Gruhl, *et al.*, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on World Wide Web (WWW)*, New York, NY, 2004, pp. 491 - 501.

[7] T. Furukawa, *et al.*, "Analyzing reading behavior by blog mining," in *Proceedings of the 22nd International conference on Artificial intelligence*, Vancouver, BC, Canada, 2007, pp. 1353-1358

[8] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: divided they blog," in *Proceedings of the 3rd Intl. Workshop on Link discovery*, Chicago, IL, 2005, pp. 36-43.

[9] ExpertSystem. (Aug 1st 2010). Available: http://www.expertsystem.net/

[10] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-512, Oct 1999.

[11] M. E. J. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, p. 13, Feb 2003.

[12] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, pp. 224-227, 1979.

[13] N. Agarwal, *et al.*, "Identifying the influential bloggers in a community," in *Proceedings of the International Conference on Web Search and Web Data Mining*, Palo Alto, CA, 2008, pp. 207-218.