Blogging Together: An Examination of Group Blogs

Marti A. Hearst
School of Information, UC Berkeley
102 South Hall, Berkeley, CA 94720
hearst@ischool.berkeley.edu

Susan T. Dumais
Microsoft Research
One Microsoft way, Redmond, WA, 98052
sdumais@microsoft.com

Abstract

This paper presents the first examination of the growing phenomenon of multi-author, or group blogs. Using a large collection of blog posts gathered over an eight month period, we conducted some quantitative assessments. Multi-author blogs were found to differ significantly from single-and-two author blogs on several quality metrics, including having longer posts on average and being ranked higher according to an in-link metric. Visualization of patterns of posting revealed that individual authors often dominate the posts in multi-author blogs, and that in some cases tags are shared, but usually not extensively.

Introduction

Blogs have captured the imagination of the public as well as researchers and are an increasingly important form of dialogue in the public sphere. A Pew research report (Rainie 2005) found that 7% of Internet users (approximately 8.4M people) had created a blog or web-based diary. 27% of Internet users said they read blogs (approximately 32M), and 12% had left comments or other materials on a blog.

To date, most academic work that focuses on blogs assumes they are of the single-author variety, which is consistent with their origins. However, over the last 5 years or so, multi-author blogs have obtained increasing influence, surpassing most single-author blogs in terms of popularity. Thus this paper takes an initial foray into understanding some characteristics of group blogs.

Motivation

Our initial interest was sparked by academic group blogs such as *The Language Log* and *The Database Column*, where an exclusive group of scholars chooses to write about a topic, usually in a witty and/or highly informed manner. Posts are inspired by occurrences in the press, the blogoshere, and academic publications and meetings, but also by posts made by other authors. It is a kind of online salon, but where outsiders are welcome to read and kibbitz in the form of comments, if not post directly. This use of blogs seems to be a unique kind of media, a "bridging genre" (Herring et al. 2004) between journals, real-world salons, and community forums.

We were curious to know how prevalent and popular such blogs are, how they form, and what the posting patterns are. As a first foray into these questions, we obtained a large collection of blog data and report on results using exploratory data analysis.

Definitions and Related Work

Sifry's popular statistical summary, "The State of the Live Web" (Sifry 2007), does not address multi-author blogs. The Wikipedia entry for *weblog* at the time of writing also does not mention group blogs, only saying that blogs are usually single-authored. Little academic research mentions group blogs, but those that do are summarized here.

What is a group blog? In the course of this investigation, we encountered the following different names:

Group Blog	Company Blog
Collaborative Blog	Project Blog (Plog)
Collective Blog	Team Blog
Community Blog	

Those in the righthand column relate more to work and business blogs. An additional term, "filter blog," is used to mean blogs like BoingBoing and Slashdot in which interesting news from other sources is aggregated and commented upon by the blog authors (Krishnamurthy 2002).

The literature debates how best to define blogs. Herring et al. (2004) define them as "frequently modified web pages in which dated entries are listed in reverse chronological sequence." boyd (2007) takes issue with this kind of definition, arguing that blogging is a process of communicating and interacting online, rather than an end product. An alternative way to define blogging is in terms of why people blog. Nardi et al. (2004) list five motivations for blogging, one of which is "Blogs as Community Forum." They listed four examples of multi-author blogs: two educational communities, a grassroots poetry blogging community, and a group of students who formed a political blogging group where different members were chosen because they had differing political views. Nardi et al. point out that, like forums and chat rooms, blogs are interactive, but also allow a kind of "interaction-at-one-remove" and a "safety net" as compared to more free-format forums. They note that "Blogs combine information and modulated interactivity. Bloggers value that they can post and share their thoughts without the intensive feedback associated with other forms of communication."

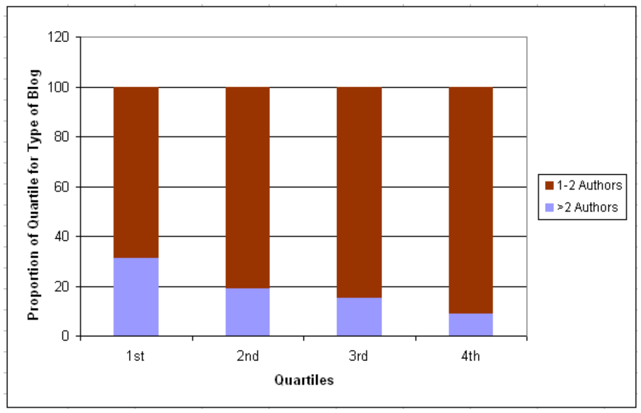

Figure 1: Bar chart showing the proportion of blogs appearing within each quartile for Technorati rankings.

People question what the boundary is between multi-authors blog and online magazines or newspapers. Baumer et al. (2008) studied how 15 people read and think about blogs. When asked "What is a blog?", the definitions varied widely, but seven referred to blogs as a newspaper or magazine, while ten used the term diary or journal, thus encompassing both ends of the spectrum. Sifry (2007)'s survey found the internet audience is becoming less likely to distinguish between mainstream media outlets and blogs.

Herrig et al. (2004), in an analysis of 203 randomly selected blogs drawn from the blo.gs site, found 90.8% were authored by one person, that 12.6% of were filter blogs, and 70.4% were personal journals. They mention Metafilter and Slashdot explicitly, stating that "[c]ommunity sites are closer to online discussion groups than are individually-maintained blogs in their frequency of activity and exchange of messages among multiple participants."

An Empirical Analysis of Multi-Author Blogs

As a first step towards understanding the nature of multi-author blogs, we examined their characteristics within a large collection of weblogs, using exploratory data analysis – a combination of database queries, visualization of aggregate statistics, and looking directly at the blogs.

We also compared group blogs and 1-2 author blogs to see if there were significant differences. We hypothesized that, on average, active multi-author blogs are of higher quality than active 1-2 author blogs. There is no automated quality measurement algorithm, but a study by Amento et al. (2000) found strong evidence linking the quality of web pages both to link-based popularity and to total number of pages on a site. One could argue that the same metrics might apply to blogs, so that blog popularity in terms of number of inlinks might be a reflection of quality.

Data Sets

Data Sources Our analysis makes use of a collection of nearly eight months of blog data collected between January 14, 2008 and August 22, 2008. The data was collected as ATOM streams produced by the SixApart company, which hosts the Blog services TypePad, LiveJournal, and Vox. LiveJournal and Vox tend to be used more as "diary" or "journal" style blogs, which we are not interested in for this study, so we used only the TypePad data. We also used heuristics to eliminate the non-English blogs, in order to facilitate interpretation. The resulting data set is referred to below as the "Raw" data collection.

The data includes the text of the posts, the date posted, the author names, and the tags assigned, but not the comments on the posts. The data also does not contain information about which blogs link to a given one, and vice versa.

Du et al. (2006) discusses the role of the software on the way blogs are used. Typepad provides a mechanism for the owner of the blog to invite other people to become authors, with some control over whether other others can place posts directly or put them in a queue to be posted. Therefore, the proportion of multi-author blogs may be higher in this blog collection than for blogs sampled in some other manner. TypePad is a hosted service, meaning that users do not have to run their own server, and it also requires users to pay a monthly fee, and so is less likely to be attractive for frivolous content than a free hosted service. Technical bloggers are more likely to host blogs themselves, and that may have effected the sampling. Finally, Typepad also supports addition of *tags*, or categories, to each post via a drop-down menu. This menu comes pre-populated with 14 tags (WebLogs, Travel, etc.) but we found in our data that most people who used tags made up their own. Sifry (2007) discusses the increasing importance of subject-oriented tags for linking information across different kinds of social media. Thus, this analysis included an examination of blogs' use of tags.

Data Characteristics The Raw dataset consists of more than 65K blogs, with a large proportion single-author (see Table 1). As to be expected, the number of blogs with increasing numbers of authors reflects a power-law curve. Upon inspection, many two author blogs were really one author using multiple aliases, so we group 1 and 2-author blogs together for the analyses below. Note that in this data, the longest time span between first post and last was 234 days, with the vast majority of blogs spanning the entire range. The mean number of posts per blog for 1-2 authors was 19.3, while for blogs with more than 2 authors it was 93.8.

In order to focus on more substantial blogs, a subset was extracted from the Raw data set that contained only those blogs with at least 33 posts, as that would be the equivalent of at least one post per week for the entire time span. This reduction serves primarily to remove low-activity single-author blogs, bringing the mean number of posts per blog for 1-2 authors blogs (112) closer to that of multi-author blogs (183). This promotes multi-author blogs to be 11% of the collection, up from 2% in the original data. This is referred to as the "Frequent" dataset.

As discussed above, we do not have a direct measure of quality, so we instead measure popularity of blogs according to Technorati rankings. Technorati is a blog data aggregation and search site; in 2007 it was tracking more than 70 million blogs. The site produces a ranking of blog popularity based on the number of blogs linking to a given blog in

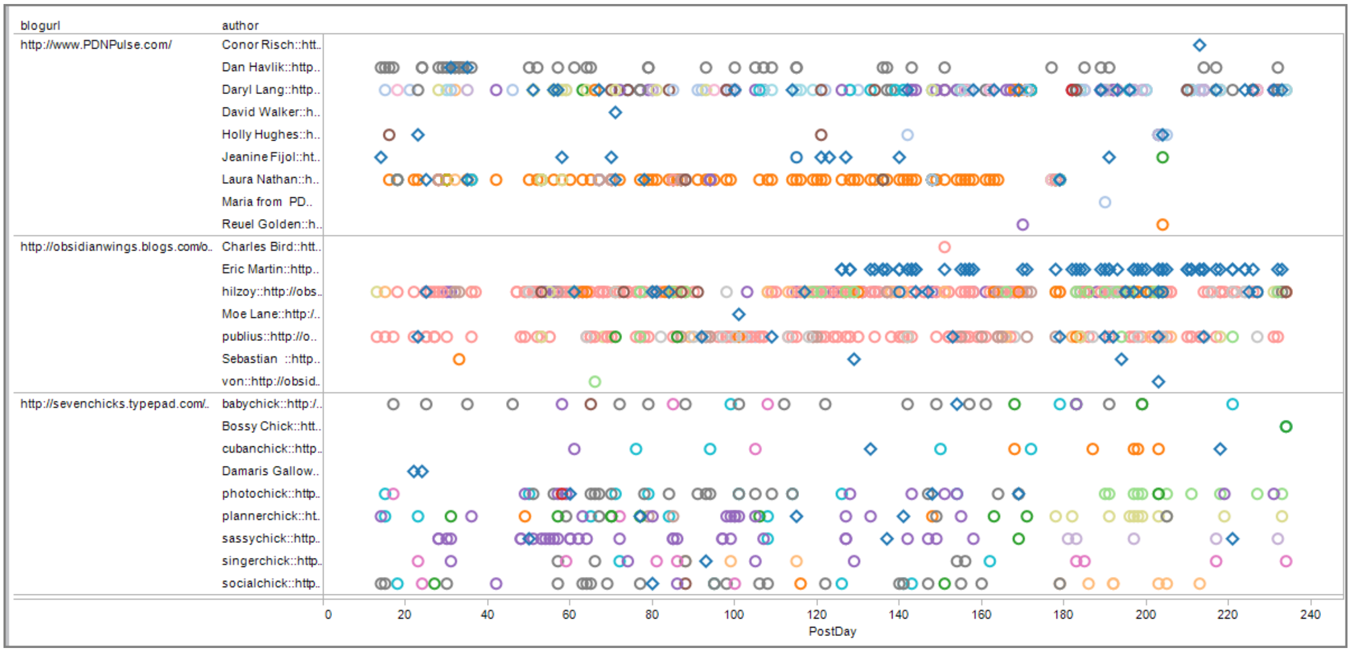

Figure 2: Posting patterns for three sample blogs, showing the authors, the number of posts, and tags assigned to the posts (diamond shape indicates no tags were used, tag text differentiated by color, only one post visible per day).

	Raw	Frequent	Technorati
Total:	65,535	9894	1061
1-2 authors:	63,700	8814	865
>2 authors:	1836 (2%)	1080 (11%)	195 (18%)

Table 1: Number of blogs in each of 3 datasets.

the last 6 months; the most linked-to blog has rank 1. This measure may be different than blog readership and of course may overlook many high-quality blogs, but can be used for rough comparisons of popularity. However, the API restricts the number of urls one can crawl, and so we use this measure on only a subset of the data. Thus, for the Technorati data set, we drew a random sample of blogs from the Frequent collection, making sure that the representation from 1-2 author blogs mirrored that of the Frequent data set (82% were 1 or 2 author blogs). For a fairer comparison, we first removed the multi-author blogs that are hosted by a major media site (wired.com, timesonline.com). This resulted in 1061 blogs. Within this set, we compared the mean number of posts per blog for 1-2 author blogs (160.3) versus multi-author blogs (188.6), and found they were not significantly different ($p < .299$), as intended. (Unless stated otherwise, we used a two-tailed t-test, assuming unequal variances.) We then obtained the Technorati rankings for this set.

Observations

Space limitations preclude a detailed discussion of qualitative aspects of the collection, but we can report that the data contained filter blogs, especially among those blogs with more than 30 authors. (For example, `palmaddicts.typepad.com` has posts about handheld electronics by 93 distinct authors.) There were also some company blogs, often used to educate users about a product or to provide more value as measured by search engines. We also observed examples of group blogs like the *Language Log* described above, although these were less common, as well as “non-blog” uses of blogging software.

Multi-author vs. 1-2 Author Blogs

We ordered the blogs in the Technorati data set by their Technorati rank, and looked at the proportion of 1-2 author blogs vs. multi-author blogs in each quartile of the ranks (see Figure 1). If popularity was not related to authorship, we would would expect to see the same proportion of multi-author blogs in each quartile (195 multi-author blogs/1061 total, or 18%). In fact, the distribution is far from uniform, with the first quartile consisting of 31% group blogs and the second having 19%. The difference was highly significant ($\chi^2(3) = 35.7, p < .01$).

Weerkam and de Rijke (2008) found that blog post length and number of comments are good predictors of information quality when assessing blog post credibility. Therefore, we compared blog post lengths for single vs. multi-author blogs. In the Technorati collection, the top ranked 50% of blogs had significantly longer average text lengths than the bottom half, ($t(1033) = 1.97, p < .0001$), independent of number of authors, thus suggesting that post length and popularity bear some relationship. Next, we compared 1-2 author blogs to multi-author blogs for average text length, finding a significant difference in favor of longer posts for multi-author blogs (2758 bytes vs. 3253, $t(249) = 1.97, p < .047$).

# authors	3	5	7	9	11	13	15
# blogs	314	141	65	52	16	26	14
AvgMax%	.73	.58	.51	.42	.36	.34	.31
Max%	.99	.98	.94	.86	.70	.79	.68

Table 2: Maximum and average maximum percentage of posts by a single author within multi-author blogs.

Thus, a first analysis using very simple metrics suggests there are significant differences in the construction of 1-2 author blogs versus multi-author blogs, and that multi-author blogs tend to exhibit more of the properties that have been found by others to be indicators of information quality.

Patterns in Author Contributions

When several authors are contributing to a blog, a natural question is how is the work load distributed: does one person pull most of the load, or perhaps two people? Or are the posts evenly distributed among the authors? Does this differ according to the number of authors? We computed, for each blog in the Frequent set, which author posted the largest percentage of posts (see Table 2). A dominant author would have a maximum percentage posts near 100%. A blog with well-distributed posts would have a maximum percent closer to the reciprocal of the number of authors, but the data shows larger than expected percentages. Even for blogs with as many as 9 authors, on average one person wrote 42% of the posts, and one author in a 9-author blog posted 86% of the time.

Figure 2 shows the detailed load sharing data for three different blogs, plotted over time, and according to which author wrote each post. Some authors dominate the conversation while others' posts are more evenly distributed. For example, the topmost blog (`www.pdnpulse.com`, a political blog) has nine authors, but two do most of the posting with a third doing an intermediate number. A frequent poster seems to drop out toward the end of the time period. Two authors post moderately often (7 and 10 times) and the remaining four authors have only one or two posts each.

Figure 2 also shows information about tag usage. Color corresponds to the content of a tag associated with the post, so the figure shows how some tags are shared among authors, and some are not. The authors for the topmost blog seem to do little sharing of tags. Two of the dominant authors tend to use primarily one tag, while another frequent poster uses a wide range of tags. The second blog (`obsidianwings.blogs.com`, also political) also has three main authors, but one began posting intensely in the middle of the time period and did not use tags. The two other dominant posters shared a number of tags. Most authors in the third blog (`sevenchicks.typepad.com`, a journaling group blog) post at about the same rate and heavily share their tags. These tags are labels with the authors names themselves, perhaps used to alert an author to look at another author's post.

One measure of community is how often a tag was shared by two or more authors within a blog. Interestingly, the average number of shared tags increases only slightly as the number of authors increases. On average, there were 3 shared tags in blogs with 3 authors, 5 shared tags in blogs with 6 authors, and 7 shared tags in blogs with 10 authors.

Conclusions

Group blogs are inherently interesting from a social media viewpoint as a hybrid way for people to interact with one another and with ideas. This paper has presented the first published research to take a systematic look at the nature of multi-author blogs. A quantitative analysis suggests that multi-author blogs, although relatively infrequent, have a significant impact on the blogosphere, being more likely in our data to be ranked higher than 1-2 author blogs according to popularity scores. They also tend to have longer average post lengths in our data, which can be an indicator of quality. These findings suggest that group blogs should receive special treatment in the design of blog search interfaces (Hearst, Hurst, and Dumais 2008). This data does not answer the questions of why people choose to form a group blog, how they choose the team members, and how they coordinate (if they coordinate) blog post writing. Interviews and surveys will most likely be necessary in order to get an understanding of these phenomena.

Acknowledgements We thank Matthew Hurst for supplying the blog data, and Dan Liebling and Steve Bush for other technical support. This research was conducted while the first author was visiting Microsoft Research.

References

Amento, B.; Terveen, L.; and Hill, W. 2000. Does ”authority” mean quality? Predicting expert quality ratings of Web documents. In *Proceedings of ACM SIGIR 2000*, 296–303.

Baumer, E.; Sueyoshi, M.; and Tomlinson, B. 2008. Exploring the role of the reader in the activity of blogging. *Proceedings of ACM CHI 2008* 1111–1120.

boyd, d. 2007. A Bloggers Blog: Exploring the Definition of a Medium. *Reconstruction* 6(2006):30.

Du, H., and Wagner, C. 2006. Weblog success: Exploring the role of technology. *International Journal of Human-Computer Studies* 64(9):789–798.

Hearst, M.; Hurst, M.; and Dumais, S. 2008. What Should Blog Search Look Like? *CIKM Workshop on Search in Social Media*.

Herring, S.; Scheidt, L.; Bonus, S.; and Wright, E. 2004. Bridging the gap: a genre analysis of weblogs. *Proceedings of HICSS* 101–111.

Krishnamurthy, S. 2002. The multidimensionality of blog conversations: The virtual enactment of September 11. *Internet Research* 3.

Nardi, B.; Schiano, D.; Gumbrecht, M.; and Swartz, L. 2004. Why we blog. *Communications of the ACM* 47(12):41–46.

Rainie, L. 2005. The State of Blogging. *Pew Internet and American Life Project* 2.

Sifry, D. 2007. The state of the live web, April 2007. *Technorati Blog* 5.

Weerkamp, W., and de Rijke, M. 2008. Credibility improves topical blog post retrieval. *Proceedings of ACL08-HLT* 923–931.