By Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins

Structure and Evolution of Blogspace

A critical look at more than one million bloggers and the individual entries of some 25,000 blogs reveals blogger demographics, friendships, and activity patterns over time.

Blogs constitute a remarkable artifact of the Web. Most people think of them as Web pages with reverse chronological sequences of dated entries, usually with sidebars of profile information and usually maintained and published with the help of a popular blog authoring tool. They tend to be quirky, highly personal, typically read by repeat visitors, and interwoven into a network of tight-knit but active communities. We refer to the collection of blogs and all their links as blogspace. By analyzing the structure and content of more than one million blogs worldwide, we've now unearthed some fascinating insights into blogger behavior.

Illustration by Gary Clement

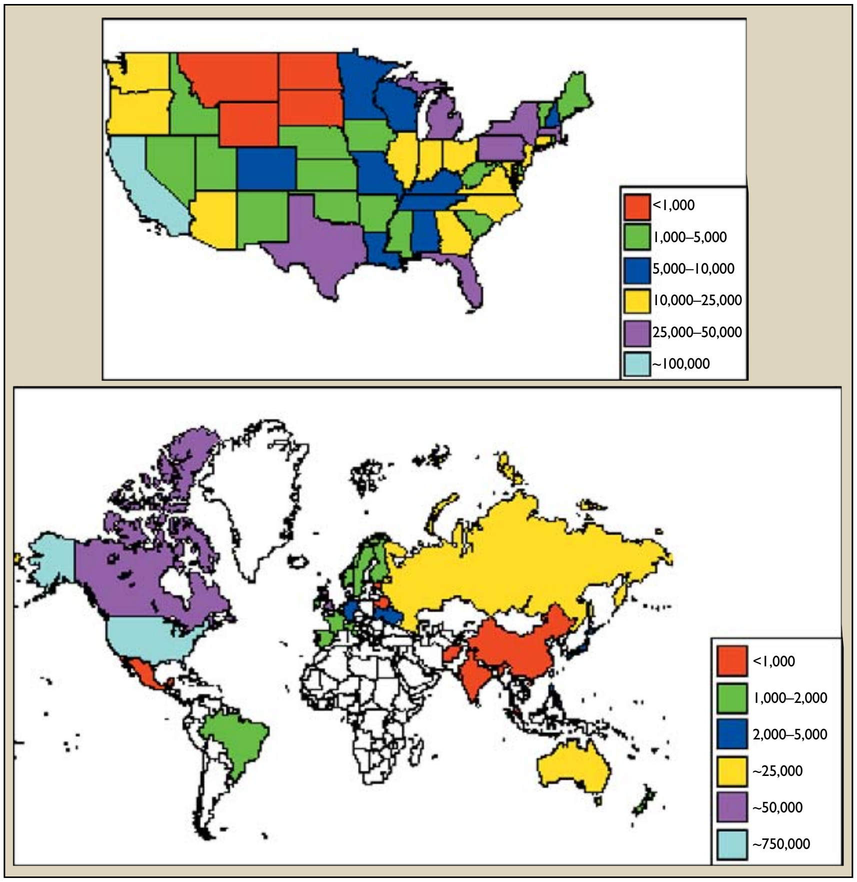

Figure 1. Worldwide geographic distribution of bloggers, Feb. 2004. (Original data source: www.livejournal.com)

An analysis of blogspace must reflect at least two distinct perspectives: the temporal (how it evolves over time) and the spatial (how bloggers congregate in terms of interests and demographics). Studying them requires data sets with distinctive characteristics; in particular, the temporal needs a time-dependent history of a collection of blogs. Here, we describe how we've studied these interests and demographics, eliciting some striking correlations in the friendships among bloggers and their interests, as well as the temporal, based on a set of blogs we've analyzed over time.

Who are these bloggers? We've studied the profile pages of 1.3 million bloggers at livejournal.com, one of the world's most popular blogging sites. Each livejournal blogger has a self-reported profile of basic personal information, including name, geographic location, date of birth, interests, friends, and other bloggers listing this blogger as a friend. Bloggers may opt not to specify or even expose certain fields; for example, only 52% of the livejournal entries (investigated February 2004) included age information. Geographic information is specified by selecting a country and optionally a U.S. state from a drop-down menu, then optionally entering an arbitrary city and an arbitrary state/province/territory for non-U.S. bloggers. Thus, geographic information is a combination of structured and unstructured data entry. Additionally, bloggers manually specify interests based on specific instructions regarding proper formulation; as a result, they share many interests, resulting in informal interest groups. There are roughly 850,000 interest groups listed at livejournal.com, 15% with only a single member. Approximately 68% of livejournal bloggers express at least one interest, with some expressing many more. Interests are wide-ranging, including, for example, vegetarianism, parenting, witchcraft, and catnip.

Where do these bloggers come from? Blogging is a global phenomenon; our data includes blogs from all seven continents, including Antarctica. But certain regions have large numbers of bloggers. As expected, centers of computing activity (such as California, Florida, New York, and Michigan) are strongly represented, as are Canada, England, Russia, and Australia. Figure 1 shows the geographic distribution of bloggers in the U.S. and worldwide.

What can be said about their ages and their interests? Table 1 lists the fraction of bloggers (whose profiles included age information) that fall into each age group, along with representative interests for each group.

Three out of four livejournal bloggers are between 16 and 24 years of age. Their interests (and friendships) are highly correlated with age. The surprising category of 1–3 year olds consists of individuals creating blogs for their pets and newborn children. The age-correlated interest groups showed a steady progression from early high school (MTV's Fuse network, rocker Adam Carson, drama club) through college (dorm life, frat parties), to 20-something lifestyle (Long Island iced tea, Liquid Television, bar hopping, grad school), into a more refined 30s (my kids, parenting, Doctor Who, and Bloom County), a somewhat conflicted 40s (Society for Creative Anachronism, Babylon 5, gardening), and even into later life (wine, cooking, travel). Many of the strongly age-correlated interests are completely unfamiliar to most people outside the age group; for us, exploring them has been a source of accelerated extracurricular learning.

Age	%	Representative Interests
1–3	0.5	treats, catnip, daddy, mommy, purring, mice, playing, napping, scratching, milk
13–15	3.5	Web designing, Jeremy Sumpter, Chris Wilson, Emma Watson, TV, Tom Felton, FUSE, Adam Carson, Guyz, Pac Sun, mall, going online
16–18	25.2	198(6, 7, 8), class of 200(4, 5), Dream Street, drama club, band trips, 16, Brave New Girl, drum major, talking on the phone, high school, Junior Reserve Officers' Training Corps
19–21	32.8	198(3, 5), class of 2003, dorm life, frat parties, college life, my tattoo, pre-med
22–24	18.7	198(1, 2), Dumbledore's army, Midori sours, Long Island iced tea, Liquid Television, bar hopping, disco house, Sam Adams, fraternity, He-Man, She-Ra
25–27	8.4	1979, Catherine Wheel, dive bars, grad school, preacher, Garth Ennis, good beer, public radio
28–30	4.4	Hal Hartley, geocaching, Camarilla, Amtgard, Tivo, Concrete Blonde, motherhood, SQL, TRON
31–33	2.4	my kids, parenting, my daughter, my wife, Bloom County, Doctor Who, geocaching, the prisoner, good eats, herbalism
34–36	1.5	Cross Stitch, Thelema, Tivo, parenting, cubs, role-playing games, bicycling, shamanism, Burning Man
37–45	1.6	SCA, Babylon 5, pagan, gardening, Star Trek, Hogwarts, Macintosh, Kate Bush, Zen, tarot
46–57	0.5	science fiction, wine, walking, travel, cooking, politics, history, poetry, jazz, writing, reading, hiking
>57	0.2	death, cheese, photography, cats, poetry

Table 1. Percentage of bloggers worldwide in different age groups and representative interests for each group. (Original data source: www.livejournal.com)

Bloggers do not express their interests randomly; certain interests tend to occur together in user profiles. By building clusters around pairs of co-occurring interests, we distilled 300 densely connected "interest clusters" (see Table 2). The first column of the table includes a label we assigned to each cluster. The second column lists representative interests from that cluster. The third and fourth columns (if reported) list the age groups and locations most strongly associated with the cluster.

Though we assigned the labels in the first column manually, the cleanliness of the clusters suggests we could do so almost automatically from the interests in the second column, then with recourse to a semantic network (such as WordNet) [2].

The locations, ages, and interests of individual bloggers paint an intriguing picture of the constituents of blogspace. But to complete this picture, we need to understand the interconnections among bloggers, or who is a friend of whom? On average, each livejournal blogger profile explicitly names 14 other bloggers as friends. In 80% of these cases, the expression of friendship is mutual; if Bob names Sally as a friend, then Sally names Bob as a friend.

Are these friendships located randomly throughout the worldwide blogger community? Are they more clustered? Researchers studying the theory of social networks calculate the "clustering coefficient" of a network of friends, defined as the chance that two of my friends are themselves friends. Previous studies [8] have typically covered much smaller networks, with clustering coefficient values ranging from 0.1 to 0.2. For our extremely large network of bloggers, the clustering coefficient is 0.2, meaning that a remarkable 20% of the time, two friends of the same blogger are themselves friends.

Cluster Label	Interests Expressed	Age Group	Location
Existentialism	Dostoevsky, Sartre, Kafka, Camus	25–29	NY
Coastal Preppies	Burberry, Diesel, Coach, Mark Jacobs, New York, Starbucks		NY, CA
Toddlers	Puppies, kitties, bunnies, sparkles, doggies	1–3	North America
New age	Zen, metaphysics, Nietzsche, quantum physics, Buddhism, philosophy, theology	25–36	Seattle
Harry Potter	Hogwarts, Slytherin, Quidditch, Ravenclaw, Gryffindor	43–48	U.K.
Coffee	Coffee, caffeine, espresso	28–30	Seattle
Outdoor activities	Kayaking, backpacking, hiking, rock climbing, mountain biking	25–36	WA
Fast food	Burger King, Wendys, Subway	16–18	FL
Vegans	Vegan, tofu, soy, PETA		
Body art	Tattoos, body art, body piercing	25–39	Australia
Russian hackers	Java, programming, Linux, php, FreeBSD, hacking, open source	22–39	Russia

Table 2. Sample "interest clusters" among bloggers worldwide. (Original data source: livejournal.com)

One possible reason for friends to be clustered so tightly is that friendships result from commonalities (such as being from the same town, being the same age, or sharing an interest in a particular topic). If two of my friends share my interest in snorkeling, they might themselves be friends due to their own shared interest. Since we know the interests, ages, and locations of the bloggers (as they've reported them), we can ask how many friendships are "explained" by these commonalities (see Figure 2). We present them as a Venn diagram to show that certain friendships might be between individuals who share an interest and are the same age.

Surprisingly, over 70% of friendships among livejournal.com bloggers can be explained by these three factors, so fewer than 30% of friendships are between bloggers of different ages, from different locations, and with no expressed shared interests. Age is the weakest explanation for friendships, while location and interest are roughly equivalent. Interest alone explains 45% of friendships, while location alone explains 55%; together these two factors explain 70%

of friendships, and 92% of friends of the same age also share an interest or location.

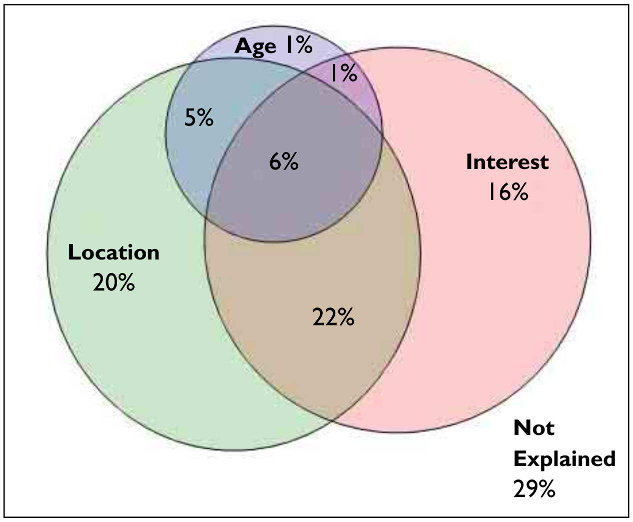

Figure 2. Explaining friendships through age, interests, and locations. (Original data source: www.livejournal.com)

Evolution of Blogspace

The culture of blogspace focuses on local community interactions among a small number of bloggers, from, say, three to 20. Members of such an informal community might list one another's blogs in a "blogroll" (a sidebar within a particular blog listing the other blogs the blogger frequents) and might read, link to, and respond to content in other community members' blogs. These sequences of responses often take place during a brief burst of activity as an interesting topic arises, jumps to prominence, then recedes. We observed and modeled this highly dynamic, temporal community structure in order to reveal the evolution of blogspace over time.

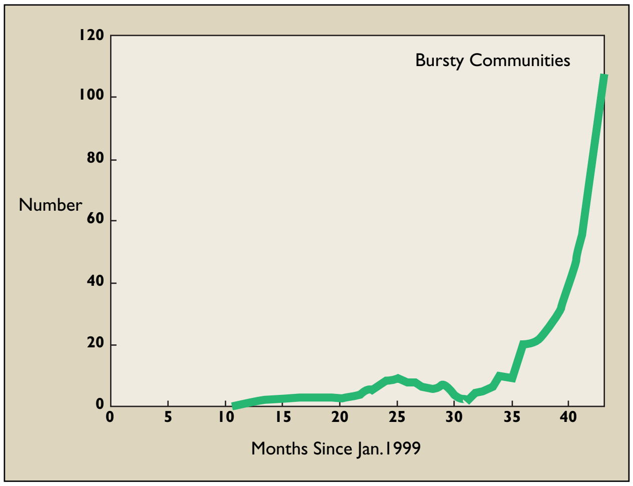

Figure 3. Burstiness of communities. (Original data sources: various blog sites)

To do so, we considered each blogger as more than a static object, extending our view of the individual blog to include a temporal component, reflecting the fact that blog entries are posted over time. It is difficult to capture the particular topics covered by each entry, as the entries lack structure; even the definition of a topic is subjective. However, we've observed that bloggers in a community often link to and cross-reference one another's postings, so we can infer community structure by analyzing the linkage patterns among blog entries.

In this view of the worldwide blogging network, a "community" is a set of blogs linking back and forth to one another's postings while discussing common topics. Each community may exhibit different levels of activity over particular periods of time; for example, a community may show a burst of rapid-fire discussion during a three-week period, then lie dormant for several more weeks before the next burst of activity.

For our study, we collected blog postings from seven popular blog sites: blogger.com, memepool.com, globeofblogs.com, metafilter.com, blogs.salon.com, blogtree.com, and Yahoo blogs. In January 2003, we crawled approximately 25,000 blogs from these sites, including all current and archived entries. Analyzing individual blogger data, we found three quarters of a million links from one of these blogs to another of these blogs, of which about 10% were distinct.

We then sought to study communities of blogs in the data to identify bursts of activity taking place in them, as illustrated in the following example. In Seattle, a group of local artists formed a blogging community around a particular blogger we call Jane. Jane was involved in fringe theater. Some of the other community members were in a band. Several events reflected the burst of activity that occurred in the community during the four months from June to October 2002. Jane decided to connect with old high-school friends, asking two members of the community to set up blogs for them. The event generated a mini-burst of blogging activity. She then convinced two high-school friends to visit Seattle on two different weekends. Lots of blogging then covered what to show them when they would visit, along with picking them up at the airport, their reaction to Jane's theater performance, and more. A third

Size	3	4	5	6	7	8	9
No.	143	165	79	14	2	1	5

Table 3. Number and size of communities in January 2003. (Original data source: various blog sites)

event during the same period occurred when two members of the community got engaged to be married, prompting another mini-bursts of blogging activity about the engagement and the beautiful children they would have.

We detected this community and the bursts within it automatically through a two-step process: extracting the communities themselves, then analyzing each community for bursts of activity.

We extracted the communities by identifying collections of blogs that frequently link back and forth to one another. Table 3 outlines the number of communities resulting from this extraction; for instance, the table reports that we found 79 communities in which a community consisted of five blogs.

Given these communities, how might social network researchers identify bursty periods of high interlinking activity? An algorithm presented in [5] identifies bursts of activity around certain words or expressions in a sequence of documents (such as an email repository). The same algorithm can be applied to the problem of finding bursts of activity in blog communities by treating each hyperlink between blogs in a community as a "word." Figure 3 shows the burstiness of communities from January 1999 to January 2003. The x axis specifies the number of months following January 1999, and the y axis specifies the number of communities worldwide displaying bursty activity.

The figure indicates an interesting pattern of behavior. The early history of blogspace—through 1999 and most of 2000—was characterized by little noticeable bursty community activity. However, there was sudden rapid growth in this activity toward the end of 2001, continuing to the beginning of 2003, the limit of the data we collected.

Interestingly, the increase in the number of bursts was not explained by the increase in the number of communities alone. Not only did the number of communities in blogspace increase over this period, the burstiness of typical communities also increased. This data suggests a change in the behavior of the bloggers themselves toward more community-oriented activity. (For more on the size and structure of blogspace, see blogcount.com; for more on the structure and dynamics of blogspace, see [1, 3, 4, 7].)

Conclusion

Blogspace is a rich and complex social environment that admits study at many levels. Our experiments are based on the profiles of more than one million livejournal.com bloggers in February 2004 and on the individual entries of some 25,000 blogs drawn from a variety of worldwide sources. A view of blogspace emerges in three layers: At the bottom is the individual blogger, who can be defined in terms of age, geography, and interests. These characteristics interact, resulting in clusters of interest groups, often with geographic or demographic correlations. In the middle is a web of friendships between pairs of bloggers. They are frequent and important and are usually explained in terms of shared locations and/or shared interests. Finally, at the top is the evolution of blog communities. They show identifiable bursts of activity that can be tracked over time. The magnitude of burstiness in communities appears to be increasing, suggesting that local community structure and community-level interactions are being reinforced as blogspace grows.

We expect blogs to remain a pervasive phenomenon on the Web, and fascinating insights into the sociology of bloggers can be divined from the analysis of the structure and content of blogspace.

References

1. Adar, E., Zhang, L., Adamic, L., and Lukose, R. Implicit structure and the dynamics of blogspace. Presented at the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference (New York, May 18, 2004); www.sims.berkeley.edu/~dmb/blogging.html.
2. Fellbaum, C. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998.
3. Gill, K. How can we measure the influence of the blogosphere? Presented at the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference (New York, May 18, 2004); faculty.washington.edu/kegill/pub/www2004_blogosphere_gill.pdf.
4. Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. Information diffusion through blogspace. In *Proceedings of the 13th International World Wide Web Conference* (New York, May 17–22). ACM Press, New York, 2004, 491–501.
5. Kleinberg, J. Bursty and hierarchical structure in streams. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Edmonton, Canada, July 23–26). ACM Press, New York, 2002, 91–101.
6. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. On the bursty evolution of blogspace. In *Proceedings of the 12th International World Wide Web Conference* (Budapest, Hungary, May 20–24). ACM Press, New York, 2003, 568–576.
7. Lin, J. and Halavais, A. Mapping the blogosphere in America. Presented at the Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference (New York, May 18, 2004); www.blogpulse.com/papers/www2004linhalavais.pdf.
8. Newman, M. The structure and function of complex networks. *SIAM Review 45,* 2 (2003).

Ravi Kumar (ravi@almaden.ibm.com) is a research staff member in the Computer Science Principles and Methodologies Department at the IBM Almaden Research Center, San Jose, CA.

Jasmine Novak (jnovak@almaden.ibm.com) is a software engineer on the WebFountain team at the IBM Almaden Research Center, San Jose, CA.

Prabhakar Raghavan (pragh@verity.com) is the chief technology officer of Verity, Inc., Sunnyvale, CA.

Andrew Tomkins (tomkins@almaden.ibm.com) is a manager in the Computer Science Principles and Methodologies Department at the IBM Almaden Research Center, San Jose, CA.