Webology, Volume 3, Number 4, December, 2006

Geographical Distribution of Blogs in the United States

Jia Lin
Ph.D., Research Associate, Waggener Edstrom Worldwide, USA. E-mail: jialin (at) gmail.com

Alex Halavais
Assistant Professor, School of Communication, Quinnipiac University, USA. E-mail: alex (at) halavais.net

Received September 16, 2006; Accepted December 15, 2006

Abstract

Blogging has diffused rapidly over the last several years in the United States, but that diffusion has not occurred evenly. In examining the distribution of 191,294 weblogs sampled in November 2003, we find that while blogging enjoys popularity throughout the U.S., bloggers appear more frequently within particular cities. This project indexes American bloggers by three-digit zip codes corresponding to their location, and identifies the demographic factors that appear to encourage blogging. We find that cities with populations that are young, urban, and more tolerant of difference are likely to host more bloggers.

Keywords

Weblog, Blog; Geography; Geocoding; United States; Zip code; City; Information retrieval

Introduction

The emergence and swift diffusion of blogging technologies has transformed weblogs from the concern of a small number of enthusiasts into a worldwide phenomenon. By October 2006, *Technorati.com* was tracking more than 55 million blogs. According to the Pew Internet and American Life Project, 8% of adult Internet users in the United States keep weblogs, and 39% read blogs (Lenhart & Fox, 2006). Personal online publishing is entering the mainstream and the "blogosphere" (the totality of hyperlinked weblogs) is becoming an indispensable part of the information environment. While the Pew study identifies the demographic characteristics of many of these bloggers, it does not provide information about where blogging is taking place, and how that affects the content of both individual weblogs and the "blogosphere" at large.

Current research on Weblogs as social phenomenon tends to examine small parts of the blogosphere, and fails to look at it from macro perspective. For example, Trammell and Keshelashvili's (2005) study on self-presentation in blogs focusses on 209 A-list blogs. Kaye (2005) surveyed 3,747 Weblog readers online in 2003 and find the motivations for using blogs include information seeking and media monitoring, convenience, personal fulfillment, political surveillance, social surveillance, and expression and affiliation. Nardi et al. (2004) conducted an ethnographic study on a small group of bloggers around Stanford, California, and find that bloggers have strong desire to connect with audience and want to enter a "unknown -social spaces to update, inform, or advise, to greet or grumble, to pontificate, confess, create, and to think" (p. 230). Lawson-Borders and Kirk (2005), while investigating the blog as a social diary, as an organizaing tool, and as civic participatory journalism, claim that a blog can be an effective tool for campaign communication. Kerbel (2005) explores Americans' civil involvement via the blog posts on "blog for America," a popular blog for Howard Dean's presidential compaign in 2004. Political blogs, especially those highly influencial A-list blogs, have been the subject of several studies related to political agendas and public opinions in the U.S. (Lin & Halavais, 2005; Adamic & Glance, 2005; Cornfield et al., 2005; Farrell, 2006).

Relatively little attention has been paid to microcontent present in the large scale of personal websites, especially those from millions of orgainary bloggers, who have only very limited number or virtually no

readers. A single personal website can easily be submerged in a mighty torrent of millions of other similar sites, and the seeming triviality of its content might well lead to it being neglected. However, the totality of this microcontent across sites represents a social and ideological landscape. The rise in popularity of personal weblogs provides a rich resource for multifaceted sociological studies. Halavais (2003) proposes that the blogosphere is analogous to the framework of the city in that it represents an information ecology. Elaborating Robert Park's literature of human environments, he notes that neighborhood-like interactions are central to commerce and political dynamics, as well as cultural diffusion, in both the blogosphere and in cities. Lin, Halavais, and Zhang's (2006) research on blog hyperlink networks within the United States finds such links to be indicators of social relationships among American cities.

Rich, unsolicited discourse and easy accessibility make weblogs a valuable alternative for both macro and micro social science research. Compared to the traditional approach, weblogs as research objects demonstrates an obvious advantage. The data is made up of spontaneous self-reports from hundreds of thousands among the population, which makes weblogs a more objective and less costly data source when compared to surveys or interviews. Moreover, the open-ended nature of these self report reduce the framing bias of the researcher, providing an abundance of data that varies according to the subjects' own interests and attention.

The blogosphere provides two layers of information: content and relationships. Bloggers, individuals who write blogs, provide semantic data in a way similar to those found in social-psychological studies. Compared to personal journals, blogs are more public: a blogger is not only keeping a record of his or her life experience and thoughts, but also trying to share this record with other individuals. Blogging technology breaks down traditional centralized authorship and makes everyone (with Internet access) a potential author and publisher. Since text is "written for a special time and place and for the reader able to fit within that time and place" (Curry, 1996, p. 185), the occurrence of decentralized publishing sources consequently re-defines the community space. An important feature of weblog technologies is that bloggers can link to other blogs in their indexes (permanent links), subscribe to certain blogs through RSS, or link to other websites as references in their periodic entries. Blog hyperlinks can be viewed as affiliation networks, where people with similar interests or close social ties interact with and refer to each other (Garton et al., 1997). Such data is conveniently available on-line and indexed by increasing numbers of search engines and portals. For example, the National Institute for Technology and Liberal Education's blog census set out to collect as many blogs as possible across the world and crawled more than 900,000 through November 2003; *weblogs.com* indexes blogs recently updated; *geourl.org* indexes weblogs with appropriate metadata according to their geographical location; and *technorati.com* tracks the links among blogs and indexes more than four million blogs.

Postulating that each blog represents an estimate of the opinions of the local population, this project takes a preliminary step toward indexing blogs to their geographic location. The process of localizing blogs and the discourse they support allows for continuing work in psychogeography and content analysis within a geographical context. Once blogs are sorted by locality, future research may provide an indication of localized sentiment and the nature of intra-local interaction. There are two objectives in this study. First, establish the location of blogs and visualize the geographical distribution of blogging in the United States. The second is to explore indication of the demographic factors that may foster adoption and support of blogging as a practice in the United States. Accordingly, there are two research questions for this study:

1. What are the patterns of the blogging distribution in the United States?
2. What are the socio-economic factors that have impact on the blogging distribution patterns?

Research Design

Data collection

Blogs represent a target moving too swiftly for definition, and we have largely avoided this definitional issue by drawing on a census of blogs that was seeded by sites that "pinged" a server indicating that they are blogs and that they have been recently updated -in other words, sites that self-identified as blogs. Nonetheless, the process by which we estimate the location of these sites requires a brief exploration of the features shared by blogs. Generally, blogs are continually updated journals, with short dated entries in reverse chronological order (Walker, 2003). These entries generally allow visitors to add their own comments. Blogs usually link to and are linked from other blogs. In addition to the regularly updated postings, most blogs also have static content that includes things like links to favorite blogs (a "blogroll"), an "about" page describing the purpose of the blog and its author, links to archived entries, and often links to hosts or software providers. Other material might include banner advertising, counters, polls, and the like.

A census of just over 950,000 blogs from November of 2003 conducted by the National Institute for Technology and Liberal Education (NITLE) was used for this study. We limited our sample to blogs hosted on generic top-level domains (.com, .org, etc.), where most of US blogs are registered, or one of the country code domains frequently used outside the country to which they were assigned (.cc, .to, .tv, etc.).

Retrieving geographical information from the blogosphere

What does it mean for a blog to be "located" somewhere? Earlier work often took the IP address of servers to be indicative of location (Zook, 2000), and for others, mapping was largely a metaphorical process (Dodge & Kitchin, 2001). Here we are concerned with where, geographically, the author implied or conveyed he or she was writing from. We are, therefore, interested in estimating the location of the blogger, rather than the blog, the latter being in some sense nowhere. Attempts have been made to extract geographic information from various classes of websites (Buyukkokten et al., 1999; Markowetz, Brinkhoff, & Seeger, 2003), but blogs tend to share a set of features that make them more amenable to such an attempt.

There is no single index or method that indicates the geographical location of a blog, even among those who opt to release geographically identifying information to the public. IP addresses can sometimes be mapped to a particular geographic location, but the physical location of the server often has nothing to do with the place of blog authorship. Domain registries can be used to check self-hosted blogs (those with their own dedicated domain name), where the registrant's address is likely the most approximate indicator of the blogger's geographical location, assuming that the blogger is the registrant. In some cases it may be that a registered address is that of a web hosting service or small blogging host. For the majority of blogs that are hosted by web hosting services, registration information is meaningless, thus alternative methods must be sought.

Some large blog-hosting services, such as Blogger, Livejournal, Diaryland, and Xanga have user profile pages with fields for the user's city, state, and country of residence. These are, naturally, contingent upon bloggers' self-reports - as are entries in the domain registries for self-hosted blogs. Geographic information is also sometimes provided by bloggers in meta-tags within the HTML; the ICBM protocol, for example, tracked by *geourl.org* and implemented by Word Press and other blogging tools, provides a way of indicating the longitude and latitude associated with a particular web page. Blogs registered with the Blogchalk index also sometimes have their geographical information available as part of a meta-tag.

When other indicators are absent, we can sometimes infer the location of a blog using explicit or implicit references to locality in the text of the page. Many bloggers, following the custom in newspapers, like to tell readers about the local weather. These weather links often reliably provide a city name or zip code. A considerable number of blogs have biography or resume pages ("about" pages) that can also provide some indication of the blogger's location. City names can also be found from links on the index page to local media, school, church or other organizations as well as links to local blogger indexes (e.g., LAblogger.com, NYCblogger.com). While a number of different approaches to mining this data were attempted, with the exception of weather information, few were generalizable enough to consistently yield useful information.

Using the above methods, a manual pilot test on 1,500 randomly selected blogs, with domains among the generic top-level domains (TLDs) and other US-related domains, successfully yielded an estimate of the geographic location of over 850 blogs, about 200 of which were based in countries outside the U.S. The success rate was higher (about 60%) for self-hosted blogs than for blogs on hosting services (about 30%). In automating this process, we prioritized indicators based on the degree to which they appeared to accurately reflect the location of each blog, based on an informal approximation by a human reader. A crawler and parser were constructed to extract each of the following indicators. They are listed in order of preference; once information was located using one of these methods, further analysis was abandoned for a given blog.

Geotags: When present, explicit meta-tags pinpointing the geographical location of a site were the most unambiguous indicator of a given site's location. After extracting the values of longitude and latitude from meta-tags, they were mapped to zip code, if located in the United States. Unfortunately, few blogs provide such meta-data; a total of 1,490 US blogs and 2,103 non-US blogs were located this way.

Local weather: City location can be inferred by weather-related links, since the more exact location the blogger provides to the weather service, the more precise weather forecast they can receive. *Weather.com* and *weatherpixie.com* are the two dominant weather services used by bloggers. In each case, there is an indicator of the geographic location in the text of the URL; namely, either a zip code or local airport code.

The tracking of links to local weather reports provided geographical information from 2,084 blogs, among which 1,112 were from the United States.

Blogchalk profile: Blogchalk represented personal information about the blogger in machine- and human-readable forms. Included among these keywords were the home city and country of the blogger. While now largely defunct, the service was still occasionally used among the blogs surveyed during this period. This source provided geographical information at the city level from 914 U.S. blogs, and home countries for 1,805 non-US blogs.

Blogger profile at hosted blogs: Blogger, Livejournal, and Diaryland, the three major blogging host services during the period, provided web pages for user profiles where users were able to list their location. Locative data obtained using this approach are shown in Table 1. Because Blogger, at the time, provided profiles only for paying users, fewer blogs with geographical information were retrieved from this service.

Table 1: Summary of result of geo information from Livejournal, Diaryland, Blogger, and Xanga

	American blogs with city information	American blogs with no city information	Non-US Blogs	Total
Livejournal	146,842	8,367	38,130	193,339
Diaryland	4,279	1,435	2,682	8,396
Blogger	4,855	1,983	7,261	14,099
Xanga	9,231	76,282	23,544	109,057
Total	165,207	88,067	71,617	324,891

Registrant address from whois data: If these four methods failed to extract geographical information, the domain registration address of blogs that were self-hosted was checked. While many domain registrars maintain public directories of registrants (traditionally retrieved using a "whois" request), recently several prominent registrars have restricted access to this data in order to reduce its use by advertisers. This did not significantly impede our data collection, but if this trend continues, it may be that domain registrations are no longer a useful indicator of location. A tool was designed to extract registrants' zip codes (for US blogs) and country names (for non-US blogs) from the registration information. This method identified geographical information from 33,610 blogs, among which 24,138 were US blogs. A manual check of 200 randomly selected blogs shows that extracting zip codes using this method was accurate in at least 98% of the cases, despite differences in how addresses were reported.

Other textual data: As noted above, few approaches to mining data on the main page or "about" pages were effective. While it was possible to seek out phrases like "I live in . . ." as well as addresses and telephone numbers, determining whether these reflect something meaningful about the author or the site is a nontrivial process. About 250 sites were flagged using a variety of methods for seeking out city names, but it was not possible to effectively code these without human involvement.

Standardizing geo information

Retrieving geographical information from different blogs, or blogs hosted by different services, leads to a variety of identification points. For self-hosted blogs we can often retrieve their street address, and sometimes a nine-digit zip code, from their domain registration. For blogs providing geographic information via meta-tags, we can also record a very precise geographical location; limited, perhaps, by the accuracy with which users determine their own longitude and latitude. But the majority of self-reported geographic information determined from the text on a site provides only the name of a city, state, or nation. Other deeply embedded information sometimes includes the telephone area code or local airport. So how can we convert such varied forms of geographic information into one standard index that can provide a useful aggregation at the city level?

The labeling of the city unit has become increasingly vague since large-scale migration following World War II, with the expansion of the city limits of urban centers, and the emergence of "second cities" that arise between big cities (Garreau, 1996). The transformation of the manufacturing economy to an information economy, especially since the 1980s, has again changed the city landscape. The decline of the traditional industrial city has been followed by the rise of the "creative city" characterized by a concentration of high-tech industries, knowledge-based production, and urban lifestyles. These areas do not necessarily surround governmental or corporate centers, as traditional industrial cities did, and they are much more fragmented and oriented toward customization and creativity (Florida, 2002).

While geographers are quick to note that zip codes have been created and used for purely practical reasons -available space for a post office or effective delivery routes, for example- these reasons also relate to the organic development of cohesive city units. Just as today's optical fiber follows the paths of century-old railways, we might consider zip codes to be an earlier form of locating communicative boundaries. Using 5-digit zip codes corresponding to streets or blocks is too narrow, and leaves us unable to assign to a particular zip code blogs that are identified solely with a city name (Weiss, 1989). Three-digit zip codes represent a geographical unit in a way that is widely used for both product marketing and political targeting strategies, a unit that it more closely related to a common sense feel for a city space.

Though we lose some specificity, the three-digit zip code allows us to locate and compare the largest number of blogs. Blogs that identify themselves as coming from "Chicago" would otherwise not be able to be included because of the lack of specificity. For purposes of analysis, it is advantageous to utilize this broad indicator. Big cities like Chicago contain a diversity of different groups, and we may overlook significant differences contained within that unit. It is still possible, depending on the application, to restrict geographic positions to greater degrees of specificity: the five and nine-digit zip codes, or even more finely grained measurements when bloggers report their longitude and latitude. However, for an understanding of the distribution of bloggers, and for "local" demographics that are somehow reflective of the city, the three-digit code provides us with the best target for categorization.

Zip codes begin with a digit from zero to nine that indicates a general region of the U.S., from zero in the northeast to nine in the west. Each subsequent digit of the zip code further divides the area, down to a five-digit indicator of a local post office. Given the name of any city or town, there are from one to 100 possible five-digit zip codes that might correspond to it. But more often than not, the first three digits are the same within a single city. The U.S. census data, for instance, only assigns five-digit zip codes with the same first three digits to one city or town, even if some large metropolises do include zip codes that extend beyond this code.

There are four different types of areas represented by three-digits zip codes. The first type is the metropolis, such as 100 for New York City (Manhattan), 900 for Los Angeles, 606 for Chicago, or 770 for downtown Houston. The second, and the most prevalent type of area, is a cluster of suburban cities and towns bordering metropolises: 201, for example, designates the western suburbs of Washington, D.C. and 750 covers suburban cities around Houston. The third type of area includes a cluster of cities that is near to a metropolitan area: 945 includes more than 30 small cities in the East San Francisco Bay, for example, and 480 is a group of cities north of suburban Detroit. Finally, a very small number of areas represented by three-digit zip codes include metropolitan areas as well as embedded cities and towns: 021 for Boston and twenty-some other towns and 481 for Ann Arbor and its surrounds, for example.

Social factors that impact on blog distribution

An early study (Lin & Halavais, 2004) in with a smaller blog samples hosted on Liverjournal and Diaryland finds that the areas that are host to high blogger density represent concentrations of what Florida (2002) refers to as the "creative class." Early adopters were identified as those who were already computer savvy (Blood, 2000), followed by young innovators, including college students and teenagers. Kaye's survey in 2003 (2005) also found that blog users tend to be young, highly educated men with high incomes. Blogging was a relatively new practice in 2003. Therefore, neighborhoods with younger populations or affluent professionals might have higher blog adoption. Finally, is adoption of blogging as a new technology influenced by natural factors? Barnett and colleagues found that seasonal factors, such as weather and hours of daylight, could influence TV viewing pattern (1991). Blogging was a typically an indoor activity that is similar with TV viewing. So do people from colder and less sunny climes blog more? Based on these influences, it is proposed that the following factors affect blogger densities: urbanization level, higher education, age distribution, income, social tolerance, and climate.

Measures related to these factors are presented in data provided by the U.S. census data in 2000. The raw population of a zip code provides some indication, though differences in zip code size make this an unreliable indicator. The census provides a report of urban households with more direct indication. Concentration of professionals is measured by the percentage of populations with occupations in business management, financial investments, law, software and IT, science, art and higher education. Social heterogeneity, a direct indicator of the city's tolerance level, is measured by the proportion of the population identified as white, as well as the proportion of gay households; both of which are used by Florida to indicate the magnitude of diversity in a city (2002). Finally, the proportion of individuals who are identified as "single" within a region may provide an indication of regions that attract those who prefer broader social networks and un-attached community to traditional family ties.

Data analysis

All blogs with geographical information retrieved are indexed to the corresponding 3-digit zip codes. The number of blogs on each 3-digit zip code are calculated and plotted. Density of bloggers in each area will be calculated by deviding the blog numbers by corresponding populations.

To examine the socio-economic factors that have impact on the blog distribution, correlation of blog density and all hypothesized factors will be calculated. When using these variables to build a multivariate regression model that predicts blogger density, we use the logarithmic values of variables, to better represent the distribution of urban populations and effects. Stepwise linear regression generates the regression models with all independent variables significant at a .05 level.

Findings

Distribution of Blogs

Using a variety of sources of geographic information, we identified the location of 272,523 blogs (or about 26% of the total), 191,294 of which were located within the United States. At least 80% of these blogs were hosted by blog-hosting sites, of which Blogger, Livejournal, Diaryland, and Xanga had the most users.

The total number of non-US blogs registered with generic top-level domains (.com, .net, .org, etc.) was 81,229, mainly from English-speaking countries such as Canada, Britain, and Australia.

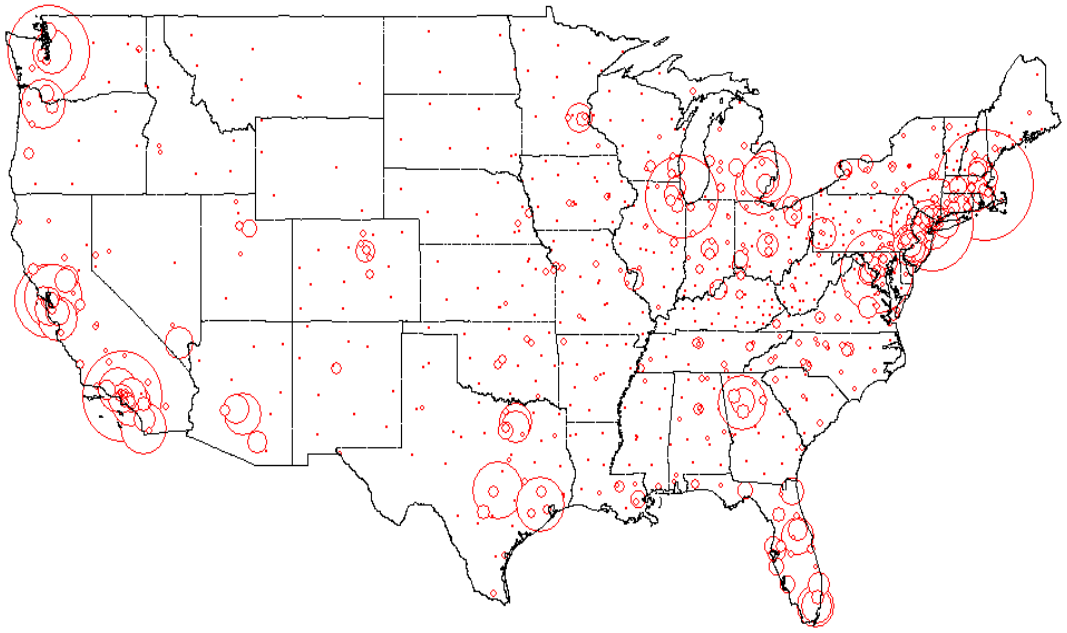
Table 2: Non-US blogs registered with generic TLDs (e.g., .com, .org, .net)

Rank	Country	NO. of blogs
1	Canada	14477
2	United Kingdom	9109
3	Australia	6173
4	Russia	4461
5	France	4266
6	Hong Kong	3701
7	Singapore	3012
8	Brazil	2477
9	Taiwan	2134
10	Japan	1841
11	Israel	1769
12	Philippine	1641
13	South Korea	1379
14	Malaysia	1308
15	Indonesia	1276
16	Portugue	1188
17	Germany	1172
18	Netherland	1085
19	Denmark	1052
20	China	803
21	Italy	631
22	New Zealand	538
23	Finland	491
24	Ireland	480
25	India	469
26	Switzerland	454
27	Vietnam	446
28	Thailand	369
29	Argentina	362
30	Belgium	330
	All others	12335
Total		81229

In converting the geographical information from each of the 188,533 blogs into three-digit zip codes, we obtained a total of 925 three-digit zip code units, and 166 of these units contain more than 300 blogs from the sample.

Blogs in the United States are roughly distributed with the population, with the largest number of bloggers located in Boston (021), New York (100), Los Angeles (900), Seattle (981), and Chicago (606). The geographic distribution of the blogs in this sample is presented in Figure 1.

Figure 1: Distribution of US blogs. Size of circles represents number of bloggers in from sample found within each 3-digit zip code.



The fifty zip codes with the greatest number of blogs include most large U.S. cities and their suburbs. The following patterns may be observed:

1. Bloggers are heavily concentrated in the large cities.
2. New technology and economic centers or clusters have fostered dense groups of bloggers: San Francisco Bay Area, Seattle, Austin, Houston, San Diego, Atlanta, and Portland
3. The suburbs and regions surrounding big cities such as Detroit, Washington D.C., San Francisco, Boston, Los Angeles are especially fertile areas for bloggers. In particular, the suburbs of Detroit have formed blogger populations of sizes much larger than their center city.
4. Borders and coastal areas of the U.S. contain more concentrated numbers of blogers, while rural states, especially across the plains, are home to relatively low densities of bloggers.

From these distribution patterns, we can conclude that the overall distribution of blogs is consistent with both population distribution and concentrations of high socio-economic status. There is a strong correlation (r=.775) between the number of bloggers and populations of 3-digit zip code units. The scatter plot shows obvious outliers above the regression line (Figure 2), which include Boston (021), New York City (101), San Francisco (941), Seattle (981), San Diego (921), and Austin (787), all cities or regions that have previously been identified as technopolises. Blog density in each area is obtained by dividing the number of blogs from the survey by the corresponding population. The areas with the greatest density of bloggers (see Table 3) include educational hubs, such as Berkeley, Boston, Pittsburgh and Madison, as well as new centers of technology and culture such as Seattle, Portland, Austin and Santa Monica.

Figure 2: Scatter plot of number of bloggers from sample in various 3-digit zip-codes against population in millions (r=.775, p<.01). Outliers are labeled with 3-digit zip-code.

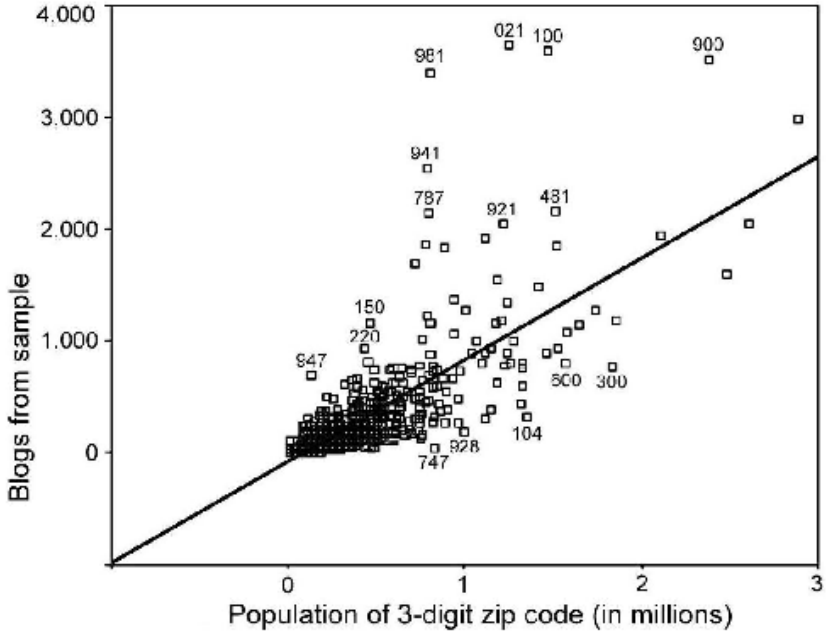


Table 3: Areas with highest blogger density

Rank	Blogs per 10K pop.	Zip code	State	Area description
1	59.27	947	CA	Berkeley
2	42.63	981	WA	Seattle
3	32.67	941	CA	San Francisco
4	32.49	943	CA	Palo Alto
5	29.71	021	MA	Boston & northern vicinity
6	27.13	787	TX	Austin
7	25.15	150	PA	Pittsburgh
8	24.80	100	NY	Manhattan
9	24.17	972	OR	Portland
10	23.44	328	FL	Orlando
11	22.64	111	NY	Long Island
12	22.02	220	VA	West of DC
13	21.71	102	NY	Manhattan
14	21.38	904	CA	Santa Monica
15	21.34	803	CO	Boulder
16	21.09	303	GA	Atlanta
17	20.17	537	NY	Albany
18	20.01	931	WI	Madison
19	19.71	122	CA	Santa Barbara
20	19.51	326	FL	Gainesville & vicinity
21	18.89	101	NY	Manhattan
22	18.80	016	MA	Worcester
23	18.29	030	NH	Southern NH
24	18.06	010	MA	Springfield
25	17.72	014	MA	Worcester
26	17.45	222	VA	Arlington
27	17.21	480	MI	North suburban Detroit
28	16.92	031	NH	Manchester
29	16.91	221	VA	Western suburbs of DC
30	16.91	921	CA	San Diego
31	16.82	041	ME	Portland
32	16.47	950	CA	Santa Cruz area

33	16.36	146	NY	Rochester
34	16.35	242	VA	Western Virginia
35	15.70	089	NJ	Princeton
36	15.36	323	FL	Tallahassee
37	15.24	900	CA	Los Angeles
38	15.18	483	MI	North suburban Detroit
39	14.95	940	CA	West Bay Area
40	14.67	276	NC	Raleigh
41	14.54	481	MI	Ann Arbor
42	14.53	126	NY	Poughkeepsie
43	14.47	358	AL	Huntsville
44	14.46	618	IL	Champaign
45	13.93	602	IL	Evanston
46	13.89	946	CA	Oakland
47	13.84	288	NC	Asheville
48	13.54	432	OH	Columbus
49	13.52	132	NY	Syracuse
50	13.30	200	DC	Washington DC

We expected the creative capacity of a region to be marked by creative occupations relating to computing and mathematics, life sciences, higher education, law, art, media, entertainment, and design. Richard Florida (2002) labels this group of professionals "the super creative core" when defining "the creative class." Annecdotally, it seems as if these professions tend to attract more bloggers.

Not each factor proposed above is highly correlated with blogger density (Table 4). The proportion of professionals and holders of college-degrees are most clearly correlated to blogging densities, while income and climate are not strong predictors. The diversity of the area, namely the proportion of non-white and gay households, is also highly correlated with blogger density. Urban population and university enrollments are positively correlated with blogger density, while the teenage population is negatively correlated with blogger density.

Table 4: Cross-tabulation of factors potentially influencing blogger density

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14
Density (1)	1													
Professionals (2)	.687*	1												
High education (3)	.641*	.882*	1											
Gay household(4)	.582*	.494*	.495*	1										
Single people (5)	.580*	.577*	.522*	.534	1									
Teenagers (6)	-.518*	-.424*	-.471*	-.376*	-.349*	1								
Urban population (7)	.513*	.505*	.558*	.429*	.568*	-.309*	1							
Univ. enrollment (8)	.455*	.510*	.470*	.240*	.683*	-.432*	.294*	1						
Income (9)	.424*	.599*	.765*	.354*	.243*	-.211*	.555*	.141	1					
Population (10)	.281*	.169	.227*	.311*	.244*	-.041	.467*	.038	.366*	1				
White population (11)	-.214*	-.228	.-161	-.449*	-.586*	-.024	-.478*	-.148	.-121	-.343*	1			
Temperature	.103	.004	-.019	.295	.017*	-.105	.255*	-.006	-.013	304	-.512*	1		

(12)														
Cloudy days (13)	-.038	-.049	-.111	-.276	-.085*	-.096	.246*	-.074	-.107	.185	.434*	-.572*	1	
Precipitation (14)	.015	.013	-.038	-.006	.047	-.299*	-.096	-.084	-.031	.021	.029	.216*	.410*	1

* correlation is significant at .05 level.

As noted, there are correlations among a number of the factors, and the degree of higher education and percentage of single people are excluded from the model to reduce collinearity. Notice that the coefficient for white population becomes positive when we control urban predictors such as professional and urban population rates. From the final models with the highest explanatory power (R^2 of .704), we can assume that in 2003, bloggers tended to concentrate in areas that were highly urbanized, liberal, densely populated, and often hubs of higher education (Table 5). Gloomy weather may contribute slightly to keep bloggers indoors and blogging. Contrary to general understanding, areas with high densities of teenagers actually have fewer bloggers.

Table 5: Stepwise multi-variate model of factors predicting blogger density

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
(Constant)	-.1354	-1.750	-1.511	-.710	-1.347	-.855	-1.604	-2.066
Professional	2.872	2.114	1.606	1.249	1.1042	1.1001	1.037	1.018
Urban population		.703	.614	.542	.617	.652	.536	.523
University enrollment			.495	.504	.542	.555	.578	.548
Gay household				.463	.613	.722	.663	.626
Cloudy days					.807	.631	.619	.591
White population						.535	.558	.525
Population							.123	.132
Teenagers								-.321
F value	689.494	543.545	426.307	359.599	335.706	301.810	271.572	239.958
R	.677	.756	.782	.799	.821	.831	.837	.839
R square	.458	.572	.611	.639	.674	.691	.701	.704

Discussion

Overall, the above analysis suggests that blogging is hardly a local phenomenon, and has a broad reach in the United States. It appears that, at least at the stage of early adoption of blogging, there is something of a "blogging divide" in America, with cities with strong creative capital hosting a larger proportion of pioneering bloggers. It may be that over time, this divide narrows.

Limitation

The procedure described in this study represents an incipient approach to inferring and estimating the geographic location of blogs. While three-digit zip code units are an acceptable indicator of city borders, there are three potential problems. One is the possible overstating of the number of bloggers in metropolises, as bloggers may list a well-known large city as theirs, rather than a less known home town. By including a region, rather than focusing on smaller suburbs as separate locations, we mitigate this overestimate somewhat. Also, when assigning 3-digit zip codes to blogs with only city names available, we used the dominant zip code for the city. Large cities such as Washington, Boston, and New York have more than one three-digit zip code. Our methods might underestimate the number of blogs in the secondary zip codes for a given city. Nonetheless, the correlation with populations does not indicate unusually low numbers in these areas.

The most difficult case is when bloggers identify themselves as coming from "Long Island" or "Orange County," which are, in fact, two broad regions consisting of several 3-digit zip codes. In the work presented here, we have assigned a certain number of blogs to each of the zip codes confined by these two identifiers, based on the proportion of population in the included areas.

Finally, there is the continuing problem of what it means for a blog to be "located" somewhere. It is likely that bloggers are among the most geographically mobile of Americans, crossing local boundaries frequently, communicating with wide networks of friends and family, and commuting (or telecommuting)

across zip codes. In some ways, what we are seeking is a mental mapping of the blogosphere, and this will necessarily always remain an estimate.

Conclusion

Through the examination of tens of thousands of geolocated blogs, this research presents some significant findings about the American blogosphere as well as about American cities. The mapping of the American blogosphere shows that bloggers are heavily concentrated in large metropolises, new technology centers, some suburbs and regions surrounding big cities such as Detroit, Washington, San Francisco, Boston, and Los Angeles. Cities and coastal areas of the United States contain more concentrated numbers of bloggers, while rural areas, especially across the plains, are home to relatively low densities of bloggers. Furthermore, the areas the greatest density of bloggers (weighted by total population) can be found in educational hubs, such as Berkeley, Boston, Pittsburgh and Madison, as well as new centers of technology and culture such as Seattle, Portland, Austin, and Santa Monica. These cities are populated with a creative workforce, single people, university students, and urbanites with higher levels of education. The idea that blogs tend to overly represent young, highly educated urbanites (McFedries, 2003; Perseus, 2003, Kaye, 2005) is supported by this study. The demographics of cities with dense blogging activities are similar to the urban culture and friendly milieu of the creative class. This research suggests that blog density can be an effective indicator of concentration of creative minds in the future.

The study undertaken here reflects the distribution of bloggers in 2003, when blogging was still a new technology for general population. As noted in the introduction, the practices of blogging shift as the technology diffuses. This study engaged early adopters of blogging, and it is natural to assume that the demographics will shift somewhat as the technology diffuses through the population, moving from the concentration on young early adopters (see McFedries, 2003). In addition to this expected shift in demographics, the nature of blogging itself has evolved quickly during the period, and there are indications that it has grown more attractive to groups that may not have been interested in blogging during the period of study.

In conclusion, this study provides an interesting map of the distribution of blogs that is consistent with population and city profiles. Though the challenge of an ever-growing blogosphere is daunting, the approach detailed here can be automated to better code the millions of blogs currently maintained, and can be expanded to include blogs hosted in countries other than the United States. Using such a distribution to further indicate regional and urban relationships and differences appears to be a promising avenue of continuing study. In particular, the concentration of bloggers in urban areas suggests that using blogs as an indicator of urban attitudes and opinions may be possible. In conclusion, by better understanding *where* people blog, this index provides the potential for gauging local knowledge and culture in a new way.

References

- Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 U.S. election: Divided they blog. *LinkKDD-2005*, Chicago, IL, Aug 21, 2005.
- Blood, R. (2000). Weblogs: A history and perspective. Retrieved April 2, 2005 from http://www.rebeccablood.net/essays/weblog_history.html
- Buyukkokten, O., Cho, J., Garcia-Molina, H., Gravano, L., & Shivakumar, N. (1999). Exploiting geographical location information of web pages. In *Proceedings of Workshop on Web Databases (WebDB)*.
- Cornfield, M., Carson, J., Kalis, A., & Simon, E. (2005). Buzz, blogs, and beyond: The Internet and the national discourse in the fall of 2004. Pew Internet and American Life report. Retrieved February 26, 2006 from http://www.pewinternet.org/ppt/BUZZ_BLOGS__BEYOND_Final05-16-05.pdf
- Curry, M.(1996). *The work in the world: Geographical practice and the written world*. MN: University of Minnesota Press.
- Dodge, M., & Kitchin, R. (2001). *Mapping cyberspace*. London: Routledge.
- Farrell, H. (2006). Politics online: Blogs, chatrooms and discussion groups in American democracy. *Political Science Quarterly, 121 (3), 517-519.*
- Florida, R. (2002). *The rise of the creative class*. NY: Basic books.
- Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, 3(1).
- Garreau, J. (1996). Civilization comes to the suburbs. *New Perspectives Quarterly*, summer: 23-25.
- Halavais, A. (2003). Netwroks and flows of content on the World Wide Web. *International Communication Association*, San Diego, 2003.
- Kaye, BK. (2005). It's a blog, blog, blog world: Users and users of weblogs. *Atlantic Journal of Communication*, 13(2), 73-95.

- Kerbe, MR., & Bloom, J.D. (2005). Blog for America and civic involvement. *The Harvard International Journal of Press/Politics*, 10(4), 3-27.
- Lawson-Borders, G., & Kirk, R. (2005). Blogs in Campaign Communication. *American Behavioral Scientist*, 49 (4), 548-559.
- Lenhart, A., & Fox, S. (2006). Bloggers: A portrait of the Internet's new storytellers. *Pew Internet and American Life*. Retrieved October 15, 2006 from http://www.pewinternet.org
- Lin, J., & Halavais, A. (2004). Mapping Bloggosphere in America. Paper presented at the *Thirteenth International World Wide Web Conference*, New York City, May.
- Lin, J., & Halavais, A. (2005). Comparison of political weblogs and mainstream media. Paper presented at the *International Communication Association 2005 General Conference*. NYC, May 2005.
- Lin, J., Halavais, A., & Zhang, B. (2006). Blog network in America: Blogs as indicators of relationships among US cities. *Connections*, (forthcoming)
- Markowetz, A., Brinkhoff, T., & Seeger, B. (2003). Exploiting the Internet as a geospatial database. *ISPRS WG IV/5 Workshop on Next Generation Geospatial Information*.
- McFedries, P. (2003). Blah, blah, blog. *IEEE Spectrum*, 40(12), 60-60.
- Nardi, B., Schiano, D., & Gumbrecht, M. (2004). Blogging as social activity, or, Would you let 900 million people read your diary? In *Proceedings of the Conference on Computer-Supported Cooperative Work*. New York: ACM Press. Pp. 222-231.
- Trammell, K.D., & Keshelashvili, A. (2005). Examining the new influencers: A self-presentation study of A-List blogs. *Journalism & Mass Communication Quarterly*, 82 (4), 968 - 982.
- Walker, J. (2003). Weblog. *Routledge Encyclopedia of Narrative*. London: Routledge. Retrieved on April 2005, from http://huminf.uib.no/~jill/archives/blog_theorising/final_version_of_weblog_definition.html
- Weiss, M. (1989). *The clustering of America*. New York: Harper & Row Publishers.
- Zook, M. (2000). The web of production: The economic geography of commercial Internet content production in the United States. *Environment and Planning A 32*, 411-426.

Bibliographic information of this paper for citing:

Lin, Jia, & Halavais, Alex (2006). "Geographical Distribution of Blogs in the United States." *Webology*, **3**(4), Article 30. Available at: http://www.webology.org/2006/v3n4/a30.html

---
