

# Mapping the Blogosphere in America

Jia Lin & Alexander Halavais

SUNY at Buffalo

School of Informatics

359 Baldy Hall

Buffalo, New York

{jlin7,halavais}@buffalo.edu

## ABSTRACT

This short paper constitutes the first phase of a long-term project focused on probing American urban culture by examining the hyperlinks and text of personal weblogs. It discusses methods of extracting geographic location information from weblogs and ways of indexing weblogs to city units. After a brief introduction to the broader research plan, the paper proposes a process to automatically extract geographic information from different weblogs. From both theoretical and practical perspectives, we will explain and justify the rationale of using 3-digit zip codes as units for comparing urban cultures. A distribution of American bloggers registered with Livejournal and Diaryland, two popular blog hosting services, will be presented to demonstrate the geocoding of the blogosphere, and to compare the distribution of these two hosts in terms of concentrations of populations and demographic profiles. Finally, we will discuss how to further improve the indexing methods.

## Keywords

Weblogs, geolocation.

## 1. INTRODUCTION

The findings discussed here represent a part of a long-term research project that will use the text and hyperlinks in personal weblogs to observe localized political agendas and opinions, as well as urban mentalities, in American cities. In addition, through examining hyperlink networks and comparing blog topics among different cities, this project will probe the interplay of geography and online communication. This study is conducted on a city level rather than at an individual level. While individual difference is important, we take the view that the sum of individuals in one city reflects the general trend at the macro level.

Little attention has been paid to microcontent presented in personal websites. A single personal website could 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 vividly and objectively depicts the social landscape and ideology at certain points of time and space. The burst of personal weblogs, regularly updated personal online journals, provides a rich resource for multifaceted sociological studies. Halavais [2] proposed that the blogosphere is analogous to the framework of the city in that it represents an ecology. Elaborating on Robert Park's [5] literature of human environments, Halavais noted that neighborhood-like interactions are central to commerce and political dynamics, as well as cultural diffusion, in both the blogosphere and in cities.

Postulating that each weblog represents an estimate of the opinions of the local population, this project takes a preliminary step toward indexing weblogs to their geographic location. Further the project will examine the social links and topics in each area. The index not only enables analysis for the purpose of this study, but can also serve as a valuable indicator of consumer behavior, media ratings, diffusion and innovation, as well as traffic and tourism flow.

## 2. GEOGRAPHIC LOCATION

Instead of sampling a limited number of weblogs, which would represent only a partial "blogosphere," this larger project will examine NITLE census data of roughly a half-million weblogs. Of these, more than half are registered in the U.S. and are checked for their geographic locations. About 80% of these blogs were hosted by blog-hosting sites, of which Blogger, Livejournal and Diaryland had the most users. Since a self-hosted blog website often has several web pages linking to each other, and the NITLE census crawls by page instead of site, a self hosted blog is often counted several times if it has several pages. Therefore, despite the high percentage of weblogs seemingly hosted on central servers, because of the duplicates, this most likely underestimates the true number.

There is no single index or method that could sort out the geographic location of bloggers, even among those who opt to release their identifying information to the public. 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 geographic location. However, for the majority of blogs that are hosted by blog hosting services, more sophisticated methods must be employed. Blog-hosting services such as Livejournal and Diaryland have a user profile field with information on 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 weblogs. Some bloggers also register their sites with regional or national blog indexes (e.g., lablogger.com, NYCblogger.com). Many weblogs have biography or resume pages ("about" pages) that provide the blogger's location. City names can also be found from links on the index page to local weather, school, church or other communities. Keyword searches and manual checks will be used together to achieve a higher extraction rate.

Using the above methods, a pilot test (manually) on 1,500 randomly selected blogs, with top-level domains that are US or generic (.com, etc.), successfully identifies the geographic location of a little more than 850 blogs, among which about 200

are from countries outside the US. The success rate is higher (about 60%) for self-hosted blogs than for blogs on hosting services (about 30%). We will work on an algorithm to realize automatic extraction of geographic information from a majority of blogs. At present, the procedure indicated in Figure 1 is followed. During the “text mining at index page” step, text on the designated area is checked against an atlas with the names of cities, states and nations. We are presently developing a set of scripts that automatically extract geographic information from various sources, including Whois data, GeoURL, and Livejournal and Diaryland user profiles.

**Figure 1. Extracting geographic data**

1. GeoURL metadata, if available
2. Whois query for unrecognized domains
3. Profile information when available from hosting service
4. Blogchalking
5. Text on main page
  - a. Bio / about / resume (city / zip / phone)
  - b. Regionalized links (weather, media)

### 3. STANDARDIZING GEO INFORMATION

Retrieving geographic information from different blogs, or blogs hosted by different services, leads to a variety of identification points, as demonstrated in earlier attempts to map the web [4]. For self-hosted blogs we can sometimes retrieve their precise street address, and sometimes a nine-digit zip code from domain registrations. For blogs providing geographic information via a metatag (the ICBM tag) we can also record the exact geographic location. But the majority of self-reported geographic information mined from an index page provides only the name of city, state, or nation. Other deeper embedded information sometimes includes the telephone area code or local weather. The question becomes: How can we convert such varied forms of geographic information into one standard index that can fulfill our research mission?

In reality, the labeling of the city unit has become increasingly vague since large-scale migration following World War II, the expansion of the city limits of urban centers, and the emergence of “second cities” that arise between big cities [6]. The transformation of the manufacturing economy to an information economy 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 comfortable lifestyles. These areas do not necessarily surround governmental or corporate centers, as traditional industrial cities did, and they are much more fragmented and oriented to customization and creativity [1]. According to Kotkin [3], the majority of middle class families lived and worked in what he terms *Midopolis*, or the old suburbs. *Nerdistans* have grown up around the fringes of *Midopolis* and provide an escape from crowding and urbanism of *Midopolis*; since the early 1990s, there have been skilled professionals migrating from urban areas to the hinterlands. Central cities have either declined into recession or undergone a renaissance to transform into new centers of culture and entertainment. In a nutshell, simple division of central, urban,

suburban and rural areas or division of cosmopolitan and provincial areas could not even closely depict the constitution of population and city types today.

So the question boils down to how to define an urban unit that matches two standards and one practical need: an ecological social system, a unit covering the newly emerged communities at the outskirts of, and even far away from, metropolitan cities, and an index indicating city units that are differentiated enough and comparable with each other? We wish to use geographic clusters consisting of certain sizes of population sharing physical proximity. Park [5] notes residential homogeneity as an important indicator in sorting neighborhoods within a city space. Rather than using city names, which are too general, or 5-digit zip codes corresponding to streets or blocks, which overly specific, 3-digit zip code units represent a middle path that defines a geographic unit in a way that is widely used in marketing and political targeting strategies.

From a practical perspective, three-digit zip codes make up a standard index to categorize blogs in geographic groups. Zip codes begin with a digit from 0 to 9 that indicates a general region of the US, from 0 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. There is rarely a one-to-one correspondence between a set of zip codes and a metropolitan area, but more often than not, the first three digits are found in the same in a single city. The US census data only assigns five-digit zip codes with the same first three digits to one city or town, though big metropolitan cities do have a few zip codes with different first three digit numbers.

The first three digit code generally represents four types of areas. First, there are three-digit codes for a metropolitan city, like 100 for New York City (Manhattan), 900 for Los Angeles, or 770 for Houston. Second, and most prevalent, is a cluster of suburban cities and towns surrounding a metropolis: zip 013, for example, designates the southern suburbs of Washington, DC and 750 covers suburban cities around Houston. The third type is a cluster of cities that is not immediately adjacent to a metropolitan area, and sometimes surrounds a mid-sized city: zip 945 includes more than 30 small cities at the East side of San Francisco Bay, and 412 is a group of cities north of suburban Detroit. Finally a very few 3-digit numbers include metropolitan areas plus embedded cities and towns, like 021 for Boston and twenty-some other towns, and 481 for Ann Arbor and its surrounds.

### 4. 3-DIGIT ZIP CODES

A distribution of American Bloggers is examined using blog data retrieved from Livejournal and Diaryland. These two blog services have become especially popular among young people and are growing rapidly. According to NITLE survey, Livejournal users increased from 56,628 in June 2003 to 316,642 in November 2003; while Diaryland users increased from 18,139 to 59,292 during the same time. Livejournal and Diaryland together feature about 28% of all blogs registered in the US through June 2003 (NITLE census data). The convenience of retrieving geographic information from the fixed user-profiles, in addition to their popularity, makes these two hosting services an ideal resource for this preliminary study. As shown in table 2, a total of 74,767 blogs from these two hosting services were crawled by the

NITLE census, and 29,259 (39.32% of total) of them provide their corresponding city names in the US.

When converting these 29,259 city names to zip codes, we obtained 797 three-digit zip code units. A breakdown of success rates for two blog hosting services is shown in table 1.

**Table1. Summary of sample blogs in Livejournal & Diaryland**

Service	US Blogs city found	US Blogs no city found	Non-US Blogs	No Geo Info.
Livejournal	24,290	15,731	7,532	9,075
Diaryland	4,969	2,005	2,661	8,504
<b>Total</b>	<b>29,259</b>	<b>17,736</b>	<b>10,193</b>	<b>17,579</b>

The distribution of Livejournal and Diaryland users is shown in table 2, and a map of their distribution in figure 2.

**Table 2. Top 50 most popular blogging areas**

Rank	# bloggers	area description & 3-digit zip code
1	531	Boston & Vicinity (021)
2	504	Metro Los Angeles (900)
3	498	Metro Seattle (981)
4	472	Metro Chicago (606)
5	413	Metro New York City (100)
6	382	East of SF Bay (945)
7	368	North suburban Detroit (480)
8	368	Ann Arbor & vicinity (481)
9	321	Metro Houston (770)
10	311	Metro San Diego (921)
11	300	Northwest of Suburban Detroit (483)
12	289	Metro San Francisco (941)
13	285	Metro Austin (787)
14	271	Orange country (926)
15	261	Metro Portland (972)
16	258	Metro Orlando (328)
17	256	Metro Atlanta (303)
18	245	Metro Philadelphia (191)
19	243	Brooklyn (112)
20	241	San Fernando (913)
21	235	Metro Miami (331)
22	219	East Vicinity of Phoenix (852)
23	199	East of Long Island (117)
24	196	West vicinity of Seattle (980)
25	187	Metro Pittsburgh (152)
26	187	East of LA (917)
27	185	Las Vegas (891)

28	174	Between Miami and Ft. Lauderdale (330)
29	170	Metro Minneapolis (554)
30	168	Suburban of Dallas (750)
31	167	Suburban Philadelphia North west (190)
32	167	Santa Cruz & Vicinity (950)
33	161	Metro Phoenix (850)
34	154	Metro Baltimore (212)
35	153	Metro Columbus (432)
36	150	Metro Dallas (752)
37	148	Vicinity of Newark (070)
38	147	Suburban Philadelphia south-east (080)
39	143	Suburban Washington D.C. (220)
40	143	Tampa (330)
41	142	Lowell and vicinity (920)
42	139	Tucson (857)
43	139	Davis and vicinity (956)
44	139	Sacramento (958)
45	138	San Jose (951)
46	135	Suburban Tampa (346)
47	134	Suburban Orlando (327)
48	134	Palm beach (334)
49	132	Buffalo & Vicinity (142)
50	131	Suburban of San Diego (920)

The top 100 list (the first half of which is found in table 2) covers most metropolitan U.S. cities and their suburbs. The following cities and areas are typical in the top list:

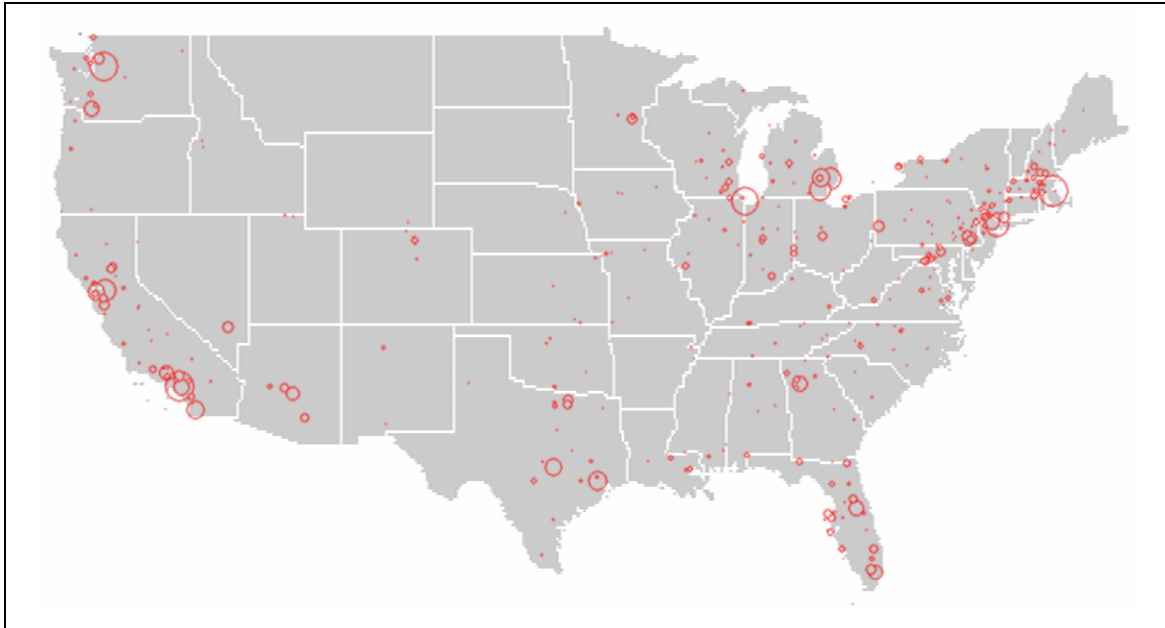
Traditional economic and cultural centers, such as Boston, New York, Los Angeles, Chicago, San Francisco, and Philadelphia.

New economic centers or clusters: Bay area, Austin, Houston, Atlanta, Orange County, East of Phoenix (Mesa, Chandler, Tempe), Las Vegas and Portland.

The suburbs and regions surrounding big cities. Regions surrounding Washington D.C. (not in top 50) and Detroit have even more bloggers than the two metropolitan areas do.

Borders and coastal areas of the US are heavily populated with bloggers, while very few inland states have concentrations of bloggers.

From these distribution patterns, we can conclude that the overall distribution is still consistent with the population distribution and concentrations of high socio-economic status. There are two concerns about using self-reported geographic information in blogs: first, a considerable number of blogs (about 31% of the total) do not provide geographic information and have to be excluded from the sample; second, the authenticity and preciseness of the geographic information may be questioned. These two uncertainties, which might affect the distribution pattern, can be reduced by this



**Figure 2. Distribution of weblogs in sample.**

pretest, demonstrating the distribution of Livejournal and Diaryland users. However, we should keep in mind the limitation of the blog resources, which are only from the two popular blog-hosting services, and might result in some bias. For example, the composition of young users might account for the higher number of bloggers in suburbs than in the city in some areas; and it might also shed light on the unusually wide dispersion and high density of bloggers in Florida. We expect a more precise density map when including other blogs, blogs that have more mature users, especially blogs that are self-hosted.

## 5. DISCUSSION

While in general, the three-digit zip code units are an acceptable indicator of city borders, there are two potential problems. The first problem is the possible overstating of the number of bloggers in metropolitan cities. Although three-digit zip codes can cover the entire metropolitan area, sometimes including its hinterlands, this provides a good estimate given that some will self-report the name of the nearest big city, rather than their own, less identifiable, home town. By including a region, rather than focusing on smaller suburbs as separate, we mitigate this estimation somewhat.

The other problem with the three-digit zip code index is in some cases lumping tens of different small cities into one suburban area and coding them as one unit. Since zip codes are designed for ease of delivery for the post office, a three-digit zip code is often assigned to a group of cities only by physical proximity to the central city, and there is no evidence of common traits or social cohesion in many of these areas. Although varied in number, Livejournal and Diaryland users are distributed in all kinds of cities. We compared Livejournal and Diaryland users in the vicinity of Boston and Lowell that are covered by zip code 021

and 018 respectively (Appendix A). The mean difference (between groups) is significant only in population size and percentage of foreigners (both are higher in Boston vicinities). But household income level, educational level, and percentage of each ethnic group have similar distribution patterns in both areas. Both of the areas have cities varied from very rich to poor, and at least one concentration of minorities (Appendix B).

While such a problem mainly emerges in several major metropolitan cities and surrounds, especially those with mixed ethnic residences, (Boston, New York City, Chicago, Detroit, San Francisco, Los Angeles), the three-digit zip code index will not help to explore the complexity and diversity in these areas. To further examine these clusters, we can divide three-digit zip codes into groups based on their socio-economic profiles. More specifically, these cities will be categorized into 6 groups based on each city's median household income (MHI) and percentage of non-Hispanic whites (% of W):

Group one:  $MHI \geq \$75000$ , and  $(\% \text{ of } W) \geq 80\%$

Group two:  $MHI \geq \$75000$ , and  $(\% \text{ of } W) < 80\%$

Group three:  $\$45000 < MHI < \$75000$ , and  $(\% \text{ of } W) \geq 80\%$

Group four:  $\$45000 < MHI < \$75000$ , and  $(\% \text{ of } W) < 80\%$

Group five:  $MHI \leq \$45000$ , and  $(\% \text{ of } W) \geq 80\%$

Group six:  $MHI \leq \$45000$ , and  $(\% \text{ of } W) < 80\%$

In this way, cities indexed in one subgroup are not only physically close to each other but also share similar SES. Once SES data for all cities are collected, we may regroup the suburban cities that are under the same three-digit zip code group, and use these subgroups as units for future analysis.

## 6. CONCLUSION

Geocoding of web pages and sites is hardly a new area of research. Nonetheless, for many such studies, discovering the location of web hosts—either through IP lookups or domain registrations (e.g., Zook [7])—has been seen as an adequate gauge of geographic location. Clearly, when it comes to weblogs, this is not adequate. The procedure described above represents an incipient approach to coding such data in a methodological way. By better understanding *where* people blog, we provide the potential for gauging local knowledge and culture in an entirely new way.

## 7. REFERENCES

- [1] Florida, R. *The Rise of the Creative Class*. Basic Books, New York, 2002.
- [2] Halavais, A. Urban Sociology and a Research Agenda for the Blogosphere. Presented at *Interconnections 4.0*, Toronto, Ontario, 2003.
- [3] Kotkin, J. *The New Geography: How the Digital Revolution is Reshaping the American Landscape*. Random House, New York, 2001..
- [4] McCurley, K.S. Geospatial Mapping and Navigation of the Web. *WWW10*, Hong Kong, May 1-5, 2001.
- [5] Park, R. The City: Suggestions for the Investigation of Human Behavior in the Urban Environment. In R. Park and E. Burgess (eds.), *The City*. University of Chicago Press, 1915.
- [6] Philips, B. *City Lights: Urban-Suburban Life in the Global Society*. Oxford University Press, New York, 1996.
- [7] Zook, M. The Web of Production: The Economic Geography of Commercial Internet Content Production in the United States. *Environment and Planning A*, vol. 32, pp. 411-426.

### Appendix A: SES data in Boston and Lowell, Mass.

			Populati on 2000	Median Household income (\$)	% of non-hispanic White population	% of Hispanic population	% of African-A merican Population	% of Asian Population	% of foreign-born population	% of population with bachlor degree or higher	crime index
(018)	Andover	1	31247	87683	90.70	1.80	.70	5.00	10.10	62.50	72.30
	Billerica	1	38981	67799	93.60	1.50	1.10	2.10	.	.	.
	Lawrence	1	72043	27983	34.10	59.70	4.90	.	30.60	10.00	500.90
	Methuen	1	43789	49627	85.80	9.60	1.30	1.20	11.20	23.00	190.40
	North	1	27202	72728	92.50	2.00	.70	3.20	8.10	50.30	45.90
	North Reading	1	13837	76962	97.00	.70	.	.	5.70	41.00	30.20
	Reading	1	23708	77059	95.80	.80	.	1.40	4.30	47.80	40.10
	Tewsbury	1	28851	68800	95.70	1.20	.70	.	5.80	25.20	129.80
	Tynsborough	1	11081	69818	94.70	1.10	.	1.60	4.80	30.60	.
	Wilmington	1	21363	70652	95.80	1.00	.	1.30	5.00	31.40	177.90
	Winchester	1	20754	94049	92.40	1.00	.70	3.60	10.80	64.90	63.40
	Dracut	1	28562	57878	94.10	1.60	.80	2.60	5.40	20.10	80.50
	Haverhill	1	58969	49833	86.30	8.80	2.40	.	6.90	23.40	277.80
	Westford	1	20754	98272	92.80	1.10	.	4.20	6.30	56.80	34.00
(021)	Chealsea	1	35080	30161	38.30	48.40	7.30	4.00	36.10	10.00	.
	Revere	1	47283	37067	79.40	9.40	2.90	3.70	21.00	13.50	300.00
	Everett	1	38037	40661	75.20	9.50	6.30	2.40	21.90	14.70	285.60
	Malden	1	56340	45654	69.60	4.60	8.20	14.00	25.70	26.20	.
	Sommerville	1	77478	46315	72.70	8.80	6.50	5.50	29.30	40.60	215.20
	Quincy	1	88025	47121	78.40	2.10	2.20	15.00	20.00	31.80	195.70
	Cambridge	1	101355	47989	64.50	7.40	11.90	9.90	25.90	65.10	310.90
	Weymouth	1	53988	51665	94.00	1.30	1.40	.50	5.40	26.00	.
	Medford	1	55765	52476	85.00	2.60	6.10	3.40	16.20	31.70	158.60
	Winthrop	1	.	53122	93.00	2.70	1.10	.	8.70	29.00	.
	Stoneham	1	22129	56605	94.00	1.80	.90	1.70	1.70	8.10	31.60
	Braintree	1	61790	61790	93.30	1.20	1.20	1.60	8.00	31.60	224.00
	Melrose	1	27134	62811	94.50	1.00	.90	.	6.10	40.00	70.40
	Arlington	1	42389	64344	89.80	1.90	1.70	1.10	14.00	52.80	.
	Brookline	1	57107	66771	78.70	3.50	2.70	11.50	26.60	76.90	160.20
	Milton	1	26062	78985	84.40	1.70	10.20	1.30	10.00	52.20	54.90
	Belmont	1	24194	80295	89.80	1.80	1.10	5.30	14.80	63.10	49.80
	Lexinton	1	30355	96825	85.10	1.40	1.10	8.60	19.60	69.10	.
	Wellesley	1	26613	113686	88.30	2.30	1.60	4.60	10.90	75.90	80.60

**Appendix B: One way ANOVA for Cities in zip code 018 and cities in zip code 021**

**ANOVA**

		Sum of Squares	df	Mean Square	F	Sig.
Median Household income	Between Groups	730879288.8	1	730879289	1.843	.184
	Within Groups	12291694515	31	396506275		
	Total	13022573804	32			
Population 2000	Between Groups	2245375746	1	2.245E+09	5.289	**0.028
	Within Groups	12736127506	30	424537584		
	Total	14981503252	31			
% of non-hispanic White population	Between Groups	416.772	1	416.772	1.912	.177
	Within Groups	6758.589	31	218.019		
	Total	7175.361	32			
% of Hispanic population	Between Groups	2.862	1	2.862	.017	.897
	Within Groups	5202.253	31	167.815		
	Total	5205.115	32			
% of African-American Population	Between Groups	37.725	1	37.725	4.136	*0.052
	Within Groups	237.160	26	9.122		
	Total	274.884	27			
% of Asian Population	Between Groups	53.512	1	53.512	3.763	*0.063
	Within Groups	355.475	25	14.219		
	Total	408.987	26			
% of foreign-born population	Between Groups	505.921	1	505.921	7.110	**0.012
	Within Groups	2134.779	30	71.159		
	Total	2640.700	31			
% of population with bachelor degree or higher	Between Groups	46.294	1	46.294	.110	.743
	Within Groups	12682.409	30	422.747		
	Total	12728.702	31			
crime index	Between Groups	4715.481	1	4715.481	.329	.572
	Within Groups	329643.870	23	14332.342		
	Total	334359.350	24			