Linking without thinking: Weblogs, readership, and online social capital formation

Cameron A. Marlow
Yahoo! Research
701 First Ave.
Sunnyvale, CA 94089
cameronm@yahoo-inc.com

June 8, 2006

Abstract

Weblogs have emerged as a popular form of online communication, driven by an array of personal, professional and social motivations. These websites provide tools for interaction, but designed as a broadcast medium, the depth and breadth of personal relationships between authors are not immediately observable.

Hypertext links made between blogs have been described as conversation, affiliation, or readership, implying a form of implicit social structure. We investigate this network of links using data collected through the automated surveillance of one million weblogs over the course of a month. These data suggest that attention in this economy is related to the author's frequency of communication.

To better understand the personal and social implications of weblog authorship, we have conducted a random survey of weblog authors. The results suggest two general classes of authors, professional and social, with differing motivations, behaviors, and effects in offline social life. While professional authors invest more time and entertain larger audiences, social bloggers tend to have more personal contact with their readers, and are more likely to have social capital embedded in ties formed online.

1 Introduction

The medium of weblogging has much in common with other forms of personal publishing that have been the basis of the web since its conception. A few years ago, only a handful of authors were creating websites identified as weblogs, but undoubtedly there were many thousands of others who updated their personal homepages nearly as frequently, and writing in a similar style. What distinguishes weblogging from previous web media is the extent to which it is *social*; the medium of blogging came into existence when these authors recognized themselves as a community.

Among weblogs, a considerable amount of social order is observable through hypertext links that authors make to each other's sites. These links come in two discernible forms: *static links* are typically made on the periphery of a site, either as a sign of readership, support, or marker of a social relationship. *Dynamic links*, on the other hand, are made when an author points to specific content on another weblog, signifying a conversation or acknowledgment of interest. Dynamic links tend to occur inline with the text of a weblog, and as the weblog is updated, they fall off of the front page; explicit links tend to remain regardless of how often the content is changed. Together these references form a *readership network* that spans the media produced by the community.

Within the academy weblogs have been described as many different social forms: conversations (Herring, Kouper, Paolillo, Scheidt, Tyworth, and Welsch, 2005), communities (Efimova and de Moor, 2005), bursty interactions (Kumar, Raghavan, Novak, and Tomkins, 2003), and political debate (Adamic and Glance, 2005). At their core, blogs are a communication tool, allowing one individual to communicate to an audience;

to this end, they could support any number of different types of social interaction. The distribution of readership and attention across the community is not uniform, but rather concentrated on a few individuals, with most blogs only having a few links (Marlow, 2004). How does authority and opinion leadership arise in this online community? Also, webloggers do not tell us what their hypertext links actually mean: do they imply friendship, weak ties, or simply acknowledge readership? Do they communicate frequently in other media, or is blogging their only form of interaction?

This paper addresses these questions through an analysis of the social behaviors of blogging, in two parts. First, we present an analysis of the global structure of links made between blogs; these data were collected through the automatic aggregation of weblog content over a one-month period. The observed network exhibits a degree distribution following a power law, with a few authors gathering the large majority of links while most weblogs only have a few. The source of this disparity is shown to be related to the amount of time invested into a weblog, seen through the frequency with which authors update their sites.

Second, to validate these observations, we present the results of a large survey of weblog authors in the areas of weblog usage, behaviors and social capital. Two primary motivations for blogging are identified, professional and social, with the latter exhibiting a larger readership and higher levels of investment. While investment into one's weblog is shown to be associated with higher social capital, social bloggers tended to be more likely to have acquired social capital through online relationships and professional bloggers through offline ties. The links made between weblogs are found to signify interest and readership more often than deeper social relationships while increasing levels of acquaintanceship are associated with more multiplex communications and recent readership.

2 Related Work

In the past few years weblogs have entered the attention of many academic disciplines, and generated public discourse around the social and cultural impact of this community. We will present research dealing with models and analyses of link topologies and social structure and review applicable work in the areas of computer mediated communication (CMC) and internet sociology.

2.1 Online communities

Early research on the social nature of the internet focused heavily on social ties formed online. In his description of the early online community known as "The Well," Rheingold showed that people without prior contact came together around mutual interests and personal interest, providing conversation, information, and social support (Rheingold, 1994). As opposed to offline ties, these relationships are often more specialized, centered around one or a few interests (Wellman and Gulia, 1999). However, online relationships do not stay online forever; with continued social interest, users tend to increase the the number of simultaneous communication media (multiplexity) of their communication and eventually meet face-to-face (Parks and Floyd, 1996).

While much of the CMC research focuses on surveys or ethnographic interviews, some methodological advancements have been made in the computational analysis of online interaction. The work in this area has typically approached data acquisition from perspective of the individual; by using pre-existing archives or by watching a person over time, large sets of personal interactions can be culled, and structural analysis tools applied to the resulting ego-networks. Since many people keep extensive email archives, these have been a popular source of social data (Haythornthwaite, 2000), with in- and out-links being determined by emails received and sent to other individuals.

Smith provides a systemic analysis of a community, looking at the conversations occurring on Usenet over a few months; in this research he has devised different measures of social exchange, a user typography, and global characteristics of the entire system (Smith, 1999). Similarly, a number of projects have attempted to infer social relationships from links on the web at large (Adamic and Adar, 2003; Gibson, Kleinberg, and Raghavan, 1998; Flake, Lawrence, and Lee Giles, 2000). While these data are much further removed from explicit social interaction, they provide perspective on the process of collecting data and allow us to start working on the hurdles posed by the analysis of large data sets.

2.2 Weblog structure

The nature of weblog interaction is quite conducive to study and has the potential to extend CMC research, since many forms of weblog affiliation are made in an explicit manner in a public forum. Bloggers' hypertext links have been seen as a network of readership and social relations in a number of different research projects (Marlow, 2003; Adar, Adamic, Zhang, and Lukose, 2004; Herring, Scheidt, Bonus, and Wright, 2004; Herring et al., 2005). Typically these studies present *static links* as a form of readership, with *dynamic links* implying discourse or interaction around a particular topic (Herring et al., 2005). Some studies extend this representation by inferring links from other features, such as content overlap and timing of updates (Adar et al., 2004; Gruhl, Liben-Nowell, Guha, and Tomkins, 2004).

Based on a subset of weblogs collected from weblog directories (such as the "weblogs" category on Yahoo!), Kumar and colleagues have looked at the whole-network properties of this community over a long period of time (Kumar et al., 2003). They extracted a sample of roughly 20,000 weblogs and a historical archive to obtain a longitudinal sample. They observed a graph of about 70k edges with dense subgraphs that revealed short, irregular periods, or "bursty" linking behavior and embedded communities that were easily extracted.

Herring et al. (2005) have recently conducted a general analytic survey of the structure of the weblog community using both quantitative and qualitative methods. Using a sample obtained from a weblog update montior, 4 random weblogs were selected, and from those weblogs a personal network of readership ties identified. This set of 5,517 weblogs was manually identified and analyzed. They found a range of different types of social interaction, from one-directional affiliation to repeated, reciprocal referencing between authors, concluding that the majority of weblogs are disconnected, while some dense cliques exist in fewer areas. Their findings suggest that contrary to the bursty nature described by Kumar et al. (2003), few weblogs actually engage in regular, reciprocal dialog.

Most of these studies have made the assumption that linking and topic similarity are in some way "social," imply "ties," but none have presented a detailed analysis of the true meaning of these relations[1]. At this point we can refer to weblog interconnections as a "readership network," but real social relations need to be empirically confirmed.

3 Methodology

Sampling and surveying weblog authors is not a straightforward task, in line with just about every other form of online social research. This section outlines the considerations around our methods for sampling weblogs, acquiring data and survey instruments.

Since there is no global, omniscient index of weblogs available, there are a number of mitigating factors that help decide which frame population to use for a weblog study. Some have used directories, such as the Yahoo! Weblog category (Kumar et al., 2003), but as the number of blogs has grown into the millions, these lists can no longer be comprehensive. Others have attempted to create their own index by crawling the web and identifying sites as blogs (Ceglowski, 2002), and while comprehensive, the task is time- and technology-intensive with low accuracy for fresh, active weblogs. This problem was an issue for many weblog applications, and addressed by a technology known as a *ping server*, systems that act as beacons for information about updated blogs. When an author updates their blog, the tool they use automatically contacts a ping server to let other applications know the location of their site. The output is list of updated weblogs along with the time they were changed.

In line with other comprehensive research on blogs (Herring et al., 2005), we have chosen to use the Blo.gs system (Winstead, 2005) which stands as the largest, public ping service on the web. The weblogs collected from this service will serve as the basis for our sample in both of the studies presented; a sample of weblogs will be collected and analyzed for global measures of the community and a random sub-sample of these sites will be employed in the random survey that follows.

3.1 Aggregator

The first part of this study is an analysis of the link structures found in the content of weblogs. In order to produce a data set for analysis, we must first collect a large corpus of weblog content; this task is executed

[1] Herring et al. (2005) have looked at this more closely, and their sample consisted of a small qualitative sample.

by a crawling and indexing system built for this study. This tool continuously monitors updates to the ping server and collects data about updated weblogs over time. The content of updated weblogs is fetched and stored; all of the external links are extracted and indexed in the event that they connote readership (i.e. links made between weblogs). These links are further distinguished as either dynamic or static based on whether they are to specific content or the front page respectively[2].

Since the weblogs that are obtained from Blo.gs are not constrained to America or even English-speaking authors, any number of languages may be used in the writing of the aggregated sites. While this should not affect the analysis, we needed to provide some facility for selecting English blogs for use in the survey section of the study. A statistical language identification system described by has been implemented to characterize the sample we have obtained from Blo.gs (Ceglowski, 2005).

The aggregator was run for a one-month period for data collection. The first stage of analysis involves a structural analysis of the readership network. Given the estimates on the total number of weblogs, the expected network will be on the order of millions of nodes, too large for many of the measures employed by social networks research. Sub-sampling this network removes the overall context, and looking only at individual nodes is intractable at this scale, but we can obtain some understanding of the global structure through simple, calculable measures.

After the data set is cleaned of obvious abnormalities, the first step is to convert it into a form that is amenable to most network measures which require connectivity. Two induced subgraphs, forming the largest connected component and the largest strongly connected components must be calculated. Given previous analysis of blog readership networks (Marlow, 2003; Adar et al., 2004), the largest connected component should account for a large percentage ($> 90\%$) of the entire network.

Previous studies of weblogs have revealed power-law distributions for both in- and out-degree of the readership network (Marlow, 2003; Kumar et al., 2003; Adar et al., 2004; Gruhl et al., 2004), and a similar result is expected here. A number of features could be related to this scaling property: frequency of posting, quality of posting, connections outside the network, and any number of demographic variables. In this part of the analysis we look to see if there is any relation between other observable variables.

3.2 Survey

The purpose of the weblog survey is to validate and extend the observations made in the previous section. The survey will be administered online over a one-month period, taking advantage of two samples: a *random sample* culled from email addresses extracted from the aggregated weblogs, requiring 5,000 addresses to achieve representivity[3], and a *self-selected sample* of those subjects who found the survey through public channels. The survey contains five sections in total, four of which we will address in this paper[4]: demographics, use of weblog technology, the meaning of links made between authors, and one on social capital.

3.2.1 Weblog Use

In the first relevant section, subjects are asked to detail their experience with weblogs along with the behaviors of their audience. To measure the level of commitment that a given subject has towards their site, the survey contains a number of questions about the time invested into various activities. Weblog acts are divided into three different pursuits: number of other sites read, number of posts per week, number of comments made on other weblogs, and a general question about the total time invested during an average week. Against these variables we measure the effect that this input had on the popularity of the subject's site, as quantified by their self-reported audience size and the comments received in an average week.

[2] Details for the construction of the crawling, indexing, and link extraction can be found in (Marlow, 2005).

[3] Given an expected target population of about 750,000 authors, a confidence level of 95% and interval of 3%, and the 20-30% response that can be expected from online surveys (Bosnjak and Tuten, 2003), these figures imply a sample of 5,330 subjects or a round 5,000 with a response rate of over 21.3%

[4] The fifth section on communication use turned out to be the most problematic; for a full description of the survey see Marlow (2005).

3.2.2 Links

After the subject submits the address of his or her weblog, the weblog's content is fetched in real time and links extracted, in the same method used by the weblog aggregator. A set of 5 links are randomly selected, and for each of these links, subjects are asked to classify the link into a number of different social categories (weblog, weblog post, or personal homepage), if applicable. Subjects are then given subsequent questions about the link based on the type specified.

The questions that follow are about the relationship between the subject and the weblog they linked to. The first question asks the type of relationship the author had with the associated blogger: friend[5], family, acquaintance, or "don't know them personally." The subject is then asked questions about this alter and their weblog: when they had last read the site, when they had last posted a comment on it, and when (if ever) they had met the author in person.

3.2.3 Social Capital

In the final section we hope to extract some information about the greater social network of the subjects. Given our survey time-constraints, the best means of gathering this information would either be a position generator (Lin and Dumin, 1986; Lin, 2001), measuring access to individuals of varying occupational prestige, or a resource generator (Van der Gaag, Snijders, and Flap, 2005; Snijders, 1999), measuring a subject's access to specific resources through their social ties.

Both of these survey instruments are established measures of social capital, and their relative accuracy is still a topic of debate. Because the resource generator phrases the questions in terms of actually acquiring resources, it naturally favored individuals and resources that are nearby, as opposed to the possible access to those resources. Since we are interested in measuring the overall size and range of a subject's weak ties in terms of both support and access to information, the natural apparatus for this section is the position generator. Additionally, the position generator has been shown to be a better choice when survey length is an issue (Van der Gaag et al., 2005).

We have adaped the instrument in Van der Gaag et al. (2005) which provides internationally standardized measures of occupational prestige in the form of ISEI socioeconomic index measures Ganzenboom and Treiman (2003). We adapted the occupation names slightly to be more recognizable to an American audience (e.g. we changed the profession *lorry driver* to *truck driver* and *estate agent* to *real-estate agent*). For each of the 30 occupations, we have asked the subject if they know such an individual and their relation to this person (acquaintance, friend or family).

We also add one new component to this instrument to evaluate the venue of tie formation, namely whether the introduction happened online or offline. By online we specify that the introduction occurred using an internet communication medium (email, instant messaging, bulletin boards, etc.), and offline as either face-to-face or on the phone. Using this variable we hope to establish any relationships between demographics, behavior and the sources of online capital that blog authors might have.

4 Analysis

During the months of May and June 2005, the weblog aggregator observed the weblog community and collected data on individual behaviors. During the second and third weeks of June, the weblog survey was presented to both a random sample of authors and also to anyone who wished to participate. This section will detail the results and analyses of these two studies.

4.1 Aggregator

The aggregator started collecting data on May 16th, 2005. Sometime during the month of May, our source of blog data, Blo.gs, was sold to Yahoo! Inc.; despite our coordinated preparation with Yahoo!, the service was unavailable between June 14th and June 16th.

Over the course of the 37 day period, over 15 million links were extracted from about 1 million weblogs. The updates observed are presented in Figure 1, showing the drop-out of the Blo.gs service towards the end

[5] We define *friend* in the same respect as Marsden (1984), or "someone you feel especially close to."

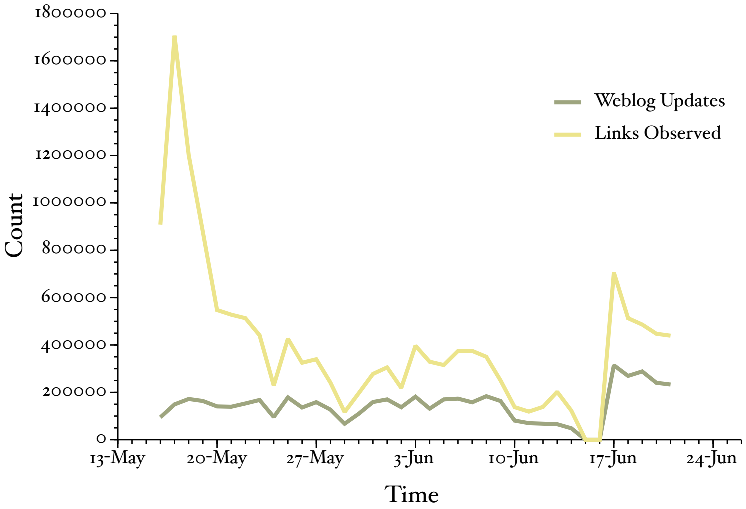

Figure 1: Weblog updates during the sample period

of the data-collection period. When weblogs are indexed initially, all links contained on the front page of the weblog are added to the database, including many that existed before the current update. This mass of relatively static links will be indexed the first time a weblog is crawled, and afterwards a much smaller set of new links will be found. This process of "getting to know" a weblog explains the severe peak and drop-off that occurs at the beginning of the data collection and shortly after the reconnection with Blo.gs.

4.1.1 Language

Of the 1,034,498 weblogs identified, 37.4% had a language detected by the classifier. English held a strong majority at 70.4% followed by Japanese with 9.9% and Spanish with 3.4%. Compared to internet market research statistics of expected Internet populations (Global Reach, 2005), the largest anomalies among this list are Portuguese and Farsi, which are far above their projected values. This data exposes some of the international biases existing in the sample obtained from Blo.gs. First, in some countries blogging is centralized around one or a few services, such as Cyworld in Korea (Lee, 2004). Since there is little need for outside aggregation of this material, these services tend not to involve themselves with ping tools. Second, some countries have their own ping services that do not interface with Blo.gs. Such ping servers are popular in Japan and France while others in Sweden, Brazil, Germany and Poland are less active.

4.1.2 Data Refinement

Because Blo.gs is an open system with a published programmatic interface, it is susceptible to a number of different types of specious activity. There are many fraudulent uses of weblogs, most of which are aimed at the individual weblogs of legitimate authors, white some involve entire weblogs. Without checking every site individually, it will be impossible to completely remove spam from this data set. However, because spam authors tend to use automated methods that create observable abnormalities, we first needed to clean the data to diminish their impact as much as possible through a number of steps of refinement.

The largest number of updates came from a weblog with over 3,000 in 34 days, or just about 88 updates per day. This amazing accomplishment suggests one of two explanations: either these updates are automated, or there is more than one person at work in changing the content of this weblog. The first method for dealing with spam is to manually check the top updated sites for fraudulent use. This technique does not cover a broad range of spammers, but it removes a large amount of inaccurate links in a short amount of time. We have deleted these weblogs, which consumed a full 85 of the top 100 updaters.

The initial readership network contained around 425,000 weblogs with at least one out-link, and about 500,000 including those with one in-links as well. Figure 2 shows the initial out-degree of the readership

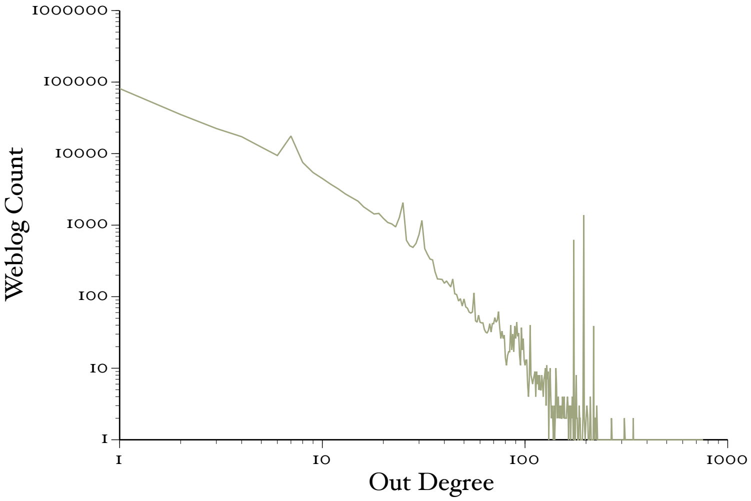

Figure 2: The initial observed cumulative out-degree distribution for the readership network.

network plotted on a log-log scale. There are a number of large spikes off of what would otherwise be a power-law distribution, most notably around the degrees of 7, 25, 31, 174, 195 and 218. Closer inspection reveals the fact that these weblogs have been automatically generated, and are not weblogs at all, but farms of spam blogs. These sites were removed from the sample by identifying regularities in their content.

By removing these weblogs from the readership network, we achieve the more believable distributions shown in Figure 3. However, the massive spike at the tail of the distribution is quite abnormal. Looking at the list of blogs with top in-degree, the first 7 blogs have 3 times the number of in-links as the next site, Slashdot, which is unmistakably one of the more popular sites on the internet. Each of these rank leaders is a weblog written by an author of the popular, open-source weblog software Wordpress (Wordpress, 2005). Their dominance in the readership network is not determined by their popularity or influence, but rather because each new installation of Wordpress comes pre-configured with links to the authors; these links were removed from the data.

4.1.3 Connectedness

To arrive at the weakly connected components, the graph is first converted to its undirected form and then searched using breadth-first-search. Starting with an initial network of 385,350 nodes and 1,970,402 edges, the largest connected component contains 343,743 nodes. Almost 90% of the weblogs updated over the sample period are in one component, a striking observation that shows the social nature of blog authorship. A qualitative inspection of the other large components ($N > 10$) reveals networks of spam, with smaller components showing blogs in foreign languages, suggesting a pocket of authors in another country who use tools that ping Blo.gs. While a path probably exists from our main component to other authors in these languages, without a global ping server, they will remain isolated.

4.1.4 Degree

The degree distribution of our readership network can be a measure of how popularity, attention, and influence is divided up amongst our blog authors. Of the links collected over the sample period, 1,399,749 static readership links were observed, and 541,234 dynamic, giving a ratio 2.6:1. Given the short time frame of the study, we had expected this ratio to be much higher, especially since, accounting for aggregation over time, our last look at similar data suggested a range of 10:1 (Marlow, 2004). The relationships between the static and dynamic in- and out-degrees are shown in Table 1.

Despite the fact that degree in these networks has an extremely high variance, we do see some relationship between the two, and in the case of In-degree, the relationship is quite strong. This association implies that

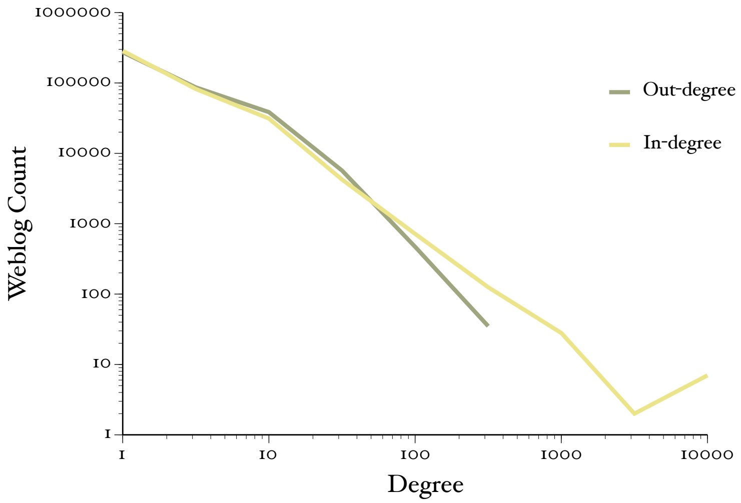

Figure 3: The initial observed cumulative degree distribution for the readership network.

	In-Static	In-Dynamic	Out-Static
In-Dynamic	0.825		
Out-Static	0.120	0.063	
Out-Dynamic	0.077	0.066	0.259

$p < 0.001$ for all measures

Table 1: Correlations between in- and out-degrees for both static and dynamic links

those blogs who are referenced statically are also being linked to with respect to specific content; without identifying the causality, it follows that those blogs who have popular content will receive links of affiliation while popular blogs will also have their writing referenced regularly.

4.1.5 Investment

What drives this popularity, either from a dynamic or static perspective? Our first assumption would be the quality of the information provided, and its general applicability to a wide range of interests. But one of the surprising characteristics of these top sites is the sheer volume of information that they produce. The top three sites across both list Slashdot, BoingBoing and Engadget, had 396, 791 and 615 updates respectively over the sample period. For BoingBoing and Engadget that amounts to over *20 posts per day*, and each from only a small number of writers.

Figure 4 shows the relationship between the number of updates made over the course of the aggregation and in-degree from dynamic links. The average number of updates over this period is shown in a cumulative fashion; for each in-degree, the value represents the average number of posts above that degree. It is also possible that a third variable, possibly writing quality, could be driving both of these features; namely, a good writer will be incented to write more as their popularity increases while ineffectual authors will be deterred by a lack of readership.

4.2 Survey

The general survey was released in two phases, first as an email to the random-sample subjects, and then publicized on a number of popular weblogs. The growth of the survey starting growing exponentially in the self-select sample in the second week after some promotional buttons were added, allowing bloggers to help advertise the survey.

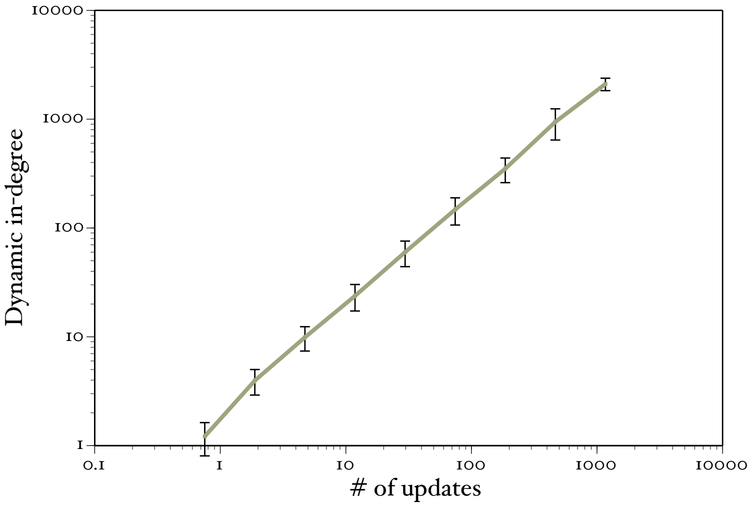

Figure 4: Updates vs. Dynamic in-degree

In the random sample, 5003 subjects were emailed at the beginning of the suvery period. Of these initial emails, 1,369 completed the survey with 3,125 not responding at all. These figures translate into a 29% response rate, very close to those obtained in other emailed random-sample surveys (Bosnjak and Tuten, 2003), a respectable response given the high probability for error in the extracted emails.

A few caveats should be made explicit. First, the survey did not fully address issues of multiple authors, which were identified in the weblog use section and removed from those sections where the survey assumed an individual author. Second, it was not available in any language other than English, despite the fact that many of the respondents live in non-english speaking countries. Finally, we did not expect the response that we received from the LiveJournal community, which accounted for about 50% of the subjects in the self-selected sample. Because the security and structure of a LiveJournal blogs is considerably different than others, it is important that we represent them as a separate sample. This partitioning gives three total sample populations: *random*, those emailed directly to participate, and two self-selected groups that found the survey through other means, *LiveJournal*, those identified as being from LiveJournal, and *self-selected*, the remaining subjects.

Table 2 contains the general demographic information (age, sex and education) for all three samples. Compared to previous broad weblog surveys (Perseus Development, 2004, 2005), the most striking difference in the demographics is the male-dominated random-sample population. A number of reasons might explain this bias, but the most likely is that men might be more willing to put their email addresses on the front page of their weblog. Comparing the LiveJournal statistics to those provided by the service (LiveJournal, 2005), the samples are extremely similar, with the exception of survey respondents being slightly older (4 years) and slightly more educated.

4.2.1 Weblog use

The first section of the survey addressed the various ways in which authors utilized their weblogs, and types of activities they engaged in on other sites including their readership, commenting, and post frequency. Table 3 shows the correlations between each of these activities for all of the samples. Nearly every investment is positively correlated with the others, suggesting that as the amount of time spent increased, so did each of these various activities. Most notable is the relationship between commenting, posting, and receiving comments. Likewise, the relationship between audience size, albeit self-reported, varies according to these investment measures.

Taking the sum of these ordinal measures, we can construct an aggregate value of investment. Figure 5 shows the relationship between this value and the self-reported audience size as a scatterplot-histogram. This relationship complements our observations of the aggregator data, showing a clear relationship between

Sample		Age	Sex	Education
Random	μ	30.2	.31	2.6
	σ	10.6	.46	1.1
	N	1,358	1,360	1,361
Self-selected	μ	29.2	.55	2.6
	σ	9.3	.50	1.1
	N	12,774	12,732	12,787
LiveJournal	μ	26.7	.71	2.4
	σ	7.5	.45	.9
	N	15,776	15,736	15,817
Total	μ	27.8	.62	2.5
	σ	8.7	.48	1.0
	N	35,254	35,195	35,327

Table 2: Sample demographics. *Age* is the number of years between the subject's birthday and the time of the survey, *Sex* is coded as 0/1 for male/female and *education* as 0 for "less than High School" with 6 as a graduate degree.

	Self-selected		LiveJournal		Random	
	Size	C_{in}	Size	C_{in}	Size	C_{in}
Read	.220	.352	.342	.542	.260	.296
Time	.271	.308	.331	.327	.319	.313
C_{out}	.475	.331	.436	.492	.459	.266
Post	.501	.322	.666	.280	.424	.324

$p < 0.001$ for all values

Table 3: Correlations between investment and audience. Investment into weblogging: *Read* is the number of weblogs an author reads weekly; *Time* is total time invested weekly, C_{out} is the frequency of comments made by the author, *Post* is their post frequency, C_{in} is the frequency of comments received, and *Size* is self-reported audience size.

investment and audience size. Unfortunately, without longitudinal data, we cannot ascertain the direction of the causality; it might be the case that the more popular weblogs inspire their authors to invest more time, or the invested time could be rewarded with larger audiences and more frequent comments. The relationship between comments posted and comments received though, regardless of its origin, suggests that commenting is not an activity that can be maintained without some investment back into the community.

4.2.2 Weblog Genres

One of the inquiries that prompted this survey was the way in which weblogs were being used. Based on the pilot data, our hypothesis was that three genres of weblogs would emerge: *journals*, with content mainly about personal experience, *editorial*, focusing on responding to news and online media, and *professional*, or a notebook for one's professional life. I assume there is quite a bit of overlap between these categories, but I expect that these will be the emerging dimensions.

In order to test this hypothesis, we took as many motivations as we could cull from interviews with pilot subjects and created one question that prompted subjects to check all of the primary reasons they kept their weblog from this list. The list of possible motivations is shown in Table 4. Each was chosen from a specific type of weblog that was given as an example by pilot subjects. Most pilot subjects classified themselves in numerous categories, with a mean of 2.5.

All ten dimensions were reduced using principal component analysis (PCA); correlations were used as

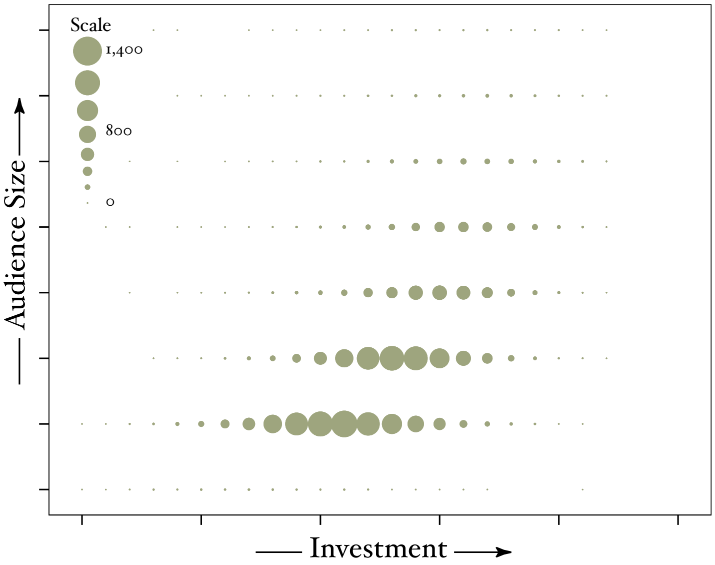

Figure 5: Relationship between author investment and self-reported audience size.

dimensions, and Quartermax rotation was applied to make the vectors easier to interpret. Two components stand out much higher than the rest; these extracted vectors are described in in Table 4.

Looking at the various contributions to these two components, it becomes clear that the first component contains all of the variables associated with professional activity while the second includes all of the purely personal motivations. Both components have some element of bookmarking and editorial writing, but in all other dimensions they are fairly divergent. Adding the third component splits the first in half, removing the editorial and linking nature from the professional component, but with only a marginal increase in the ability to describe the data.

While these motivations are not completely orthogonal, they do provide an interesting measure of why the given subject has decided to write the weblog. Our hypothesis that there were three main subjects that drove weblog—authorship, news, personal matters, and professional matters—was partially correct. It appears that news and linking are actually part of both communities described by professional and personal writing. We will refer to the components as motivational scores M_{prof} and M_{pers} respectively.

	M_{prof}	M_{soc}
Keep notes for your professional interests	.714	-.029
Increase your professional reputation	.709	-.225
To post news about an organization or project	.540	.120
Keep a list of links to things you have read	.434	.349
Comment about things you read in the news	.408	.370
Make money through advertising	.419	-.064
Keep notes or record what's going on in your life	-.171	.653
Keep in touch with friends	-.257	.644
Post photos you have taken or music you have made	.141	.608
Meet new people	.122	.550

Table 4: Results of Motivation factor analysis

4.2.3 Links

In this section, subjects were asked to answer questions about links that were extracted from their weblog during the survey, reflecting readership, social acquaintanceship and communication modalities. Because of security problems involved in accessing LiveJournal sites, very few LiveJournal subjects succeeded in answering these questions. For this reason we will only consider those subjects in the random and self-selected samples for this section.

Of the 26,075 links listed as social links, 9,700 came from a disregarded sample (LiveJournal or incomplete), and an additional 1,351 were missing data for all of the subsequent questions. After excluding these links, the responses include 15,024 samples of links listed as weblog, weblog post, or personal homepage, with 10,275 (68.3%), 2,632 (17.5%), and 1,637 (10.9%) links respectively.

Removing the personal homepages, the ratio of static to dynamic social links is about 4:1, which is smaller than the rate observed by the aggregator, 2.6:1. This discrepancy probably stems from two considerations: first, the number of dynamic links observed on all weblogs over the course of a month will aggregate to a larger number than would be found at any given day on the front pages of the same sites. Second, in cases where the subject does not know the author of a given weblog post, they might not see it as such.

		Link Type	
Relationship	Dynamic (%)	Static (%)	Homepage (%)
No Relation	69.1	55.1	43.4
Acquaintance	12.7	18.2	17.0
Friend	13.8	23.6	30.1
Family	4.4	3.1	9.5

Table 5: Social link type and relationship

Table 5 shows the associations between the various link types and the reported relationship between the subject and the other author. As would be expected, dynamic links do not necessarily imply any sort of personal interaction and static links are associated with higher levels of acquaintance than dynamic. However, the number of ties identified as having no social basis is remarkably high; over 50% were made to weblogs written by individuals with whom the subject does not even consider an acquaintance.

The next question addressed how many of these links are "live," namely weblogs that the subject reads regularly and how many are "dead," pointing to readership that no longer exists. Table 6 shows the distribution of readership as described by the last time the author read the given weblog for each type of relation to the author. While we expected to find high readership for friends' weblogs, we were surprised to see that for *all* levels of acquaintanceship over 80% of the identified weblogs had been visited in the last month, and over 60% in the last week; furthermore, the stronger the social tie described by this link, the more likely the subject is to read them regularly. Over 50% of the familial weblogs were read the day the survey was taken, and nearly 50% for those denoted as friends.

There is probably some amount of choice-supportive bias associated with affirmation of the subjects' self-image, namely believing they are much more frequent readers than they actually are. The division of time periods was explicitly chosen to minimize the generalized bias shown by the pilot subjects, but it has the downside of including large ranges of time.

Another important part of describing these social links is to determine to what extent they denote other types of social interaction. The relationships between the various social communication questions are shown in table 7. All of the measures are in terms of increasing frequency, except for friendship which is in increasing acquaintanceship.

The strongest correlations are between the level of acquaintance and various forms of communication, which should be expected. These data suggest that the stronger the tie, the more likely it is that a weblog author will use other forms of communication to interact with their weblog ties. The relationship between tie strength and the frequency of face-to-face interaction implies that in this case weblogs are part of a larger set of communication tools used to support offline ties. This is in accordance with Haythornthwaite and Wellman (1998), suggesting that the stronger a tie is, the higher the modality of interaction.

	Alter's relation to the author			
Last read	None	Acq.	Friend	Family
Never	4.5	0.9	0.7	0.8
Over a year ago	0.8	0.9	0.7	0.2
6 Mo. - 1 Year	2.0	1.8	1.2	1.7
1 Mo. - 6 Mo.	8.8	7.4	6.1	3.1
1 Week - 1 Mo.	21.5	19.4	14.0	15.2
This Week	32.7	33.8	31.6	23.6
Today	29.7	35.8	45.8	55.4
Total	100 %	100 %	100 %	100 %

Table 6: Readership and relationship

	Relat.	Read	F2F	Commented	Spoken
Relationship	1	.172	.764	.358	.770
Read	.172	1	.145	.438	.228
Face-to-face	.764	.145	1	.270	.697
Commented	.358	.438	.270	1	.419
Spoken with	.770	.228	.697	.419	1

$p < 0.001$ for all measures

Table 7: Social links and communication

Along with the observations made in the weblog use section, these data reinforce the idea that there are a number of types of interaction being expressed in the form described as a weblog. While some of the more specialized, professional weblogs can be non-social, blogging is at its core a social tool, capable of reinforcing friendships as well as allowing for new connections.

4.2.4 Social Capital

While the links section probed a sample of the subject's specific ties, we were also interested in the generalized access to resources they might have, both offline and online, even through weaker ties. The final section of the survey used a position generator to measure this form of social capital using a standardized instrument.

One of our fears about having a long survey instrument such as the position generator as the last section was that the drop-out rate in this section would be quite high. Furthermore, even though the instructions specified to check either yes or no for each position, we was also concerned that subjects would only check the "yes" answers and leave the "no" answers blank. Our apprehension proved incorrect however, as those subjects that completed the survey mostly submitted a value for all 30 occupations.

Figure 6 illustrates the responses over the entire subject population for each job, in the order that the job was presented in the survey. The gradual decline in answers suggests that as users got bored with the instrument, some percentage dropped out for every question. In analyzing the data, we will only use those subjects who completed at least 28 of the 30 questions; for those subjects who did not answer one or two questions, it is quite likely that these came as the last two questions. Since these two occupations are among the least known among the entire set, it is safe to assume that an omission of the last one or two questions is equivalent to a "no" answer. Because our measure of total social capital will be the sum of the occupational prestige scores, these left out answers should not affect the data.

I have chosen to use the sum of occupational prestige scores as our measure of social capital based on the fact that it provides the widest degree of all of possible measures. The distribution of these scores can be seen in Figure 7. The distribution shows a normal form[6] with some notable spikes towards the bottom

[6] The Anderson-Darling score for normality is 37.89 with $p < 0.001$.

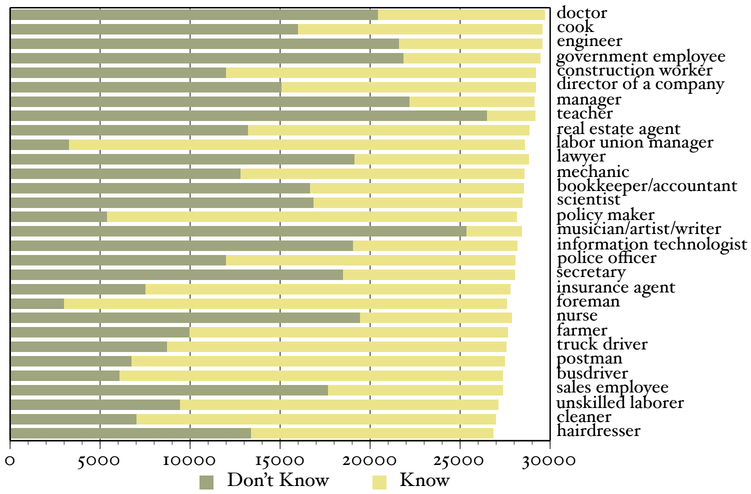

Figure 6: The number of responses per job.

and center of the distribution.

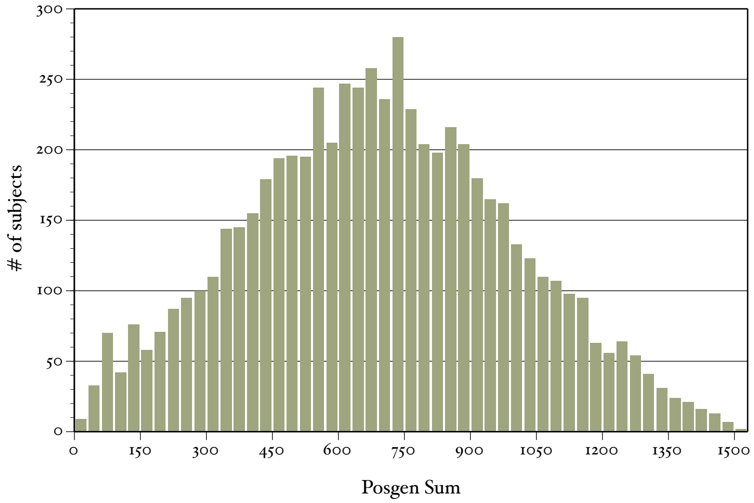

Figure 7: The distribution of position generator scores calculated as the sum of all occupational prestige scores.

The aggregate results for the position generator are presented in Table 8. Of the most common occupations, two stand out far and above the rest of the distribution; while teacher is in the top half of occupational prestige, any number of professions could satisfy this term, from pre-school to university professor. This ambiguous category makes it one of the most regularly known, and diminishes its informational value.

Likewise, the commonly known occupation of musician/artist/writer is probably caused by a misinterpretation of the question. While we was explicitly asking for people who's *profession* is the stated occupation, individuals probably interpreted this as anyone who fills that title. Since many people are amateur musicians, artists, and writers[7], most everyone knows someone who fits the description. The high rates of these two professions can explain the spikes in the lower values of the distribution; knowing only one or both of these professions places a subject in the first and second spikes respectively.

[7] Arguably, all weblog authors are amateur writers.

		prestige		*% yes*	*% if yes*				
Item #	Job	U & S	ISEI	Know	Acq.	Friend	Family	Online	Offline
11	lawyer	86	83	66	42	38	20	21	79
1	doctor	84	87	68	51	28	21	11	89
15	policy maker	82	70	*19*	55	34	11	21	79
3	engineer	76	68	72	25	45	29	17	83
17	information technologist	68	70	67	25	**59**	16	**31**	69
7	manager	67	69	76	37	46	17	16	84
6	director of a company	67	69	51	46	33	20	17	83
10	labor union manager	66	65	*11*	59	24	17	19	81
14	scientist	65	71	59	34	**51**	16	**26**	74
4	government employee	64	61	73	29	40	31	20	80
9	real estate agent	64	61	45	58	26	16	14	86
12	mechanic	63	59	44	47	28	24	11	89
8	teacher	62	66	**90**	27	48	25	18	82
18	police officer	54	50	42	54	26	20	14	86
19	secretary	52	51	65	43	42	14	20	80
19	insurance agent	52	53	27	57	25	18	16	84
13	bookkeeper/accountant	52	54	58	40	36	24	17	83
16	musician/artist/writer	45	64	**89**	20	**68**	12	**35**	65
22	nurse	44	38	69	34	33	32	16	84
26	bus driver	44	26	22	61	23	16	11	89
30	hairdresser	39	30	49	58	31	11	11	89
2	cook	39	30	53	41	41	18	17	83
23	farmer	36	43	36	38	24	38	11	89
21	foreman	27	25	*11*	48	24	28	16	84
25	postman	26	39	24	60	21	19	10	90
24	truck driver	26	34	31	44	23	33	16	84
27	sales employee	22	43	64	33	54	12	23	77
29	cleaner	20	29	25	58	28	14	13	87
38	unskilled laborer	15	26	34	42	4	16	20	80
5	construction worker	15	26	41	39	33	28	12	88

Table 8: Position generator response, associated occupational prestige and socioeconomic indicator values, and item responses

As with weak social ties in general, one assumption is that social capital will accrue as a person ages; as we get older, we meet more people, and our overall network of acquaintances grows, along with access to associated resources (Lin, 1999). Before comparing this measure of social capital to other measures in the survey, we will first look at the relationship to the demographic variables of age, gender and education.

	Age	Education	Sex
PG_{Sum}	.313	.275	-.069

N = 29,835. $p < 0.001$ for all variables

Table 9: Position Generator and demographics

We observe positive correlations between age and eduction, and a slight negative relationship with sex. Controlling for each of these variables independently does not remove these biases, so it will be necessary to control for all three in the following observations.

One of the distinguishing factors of this instance of the position generator instrument was the addition of an online vs. offline distinction for each occupation. Knowing that the typically-public nature of weblogging allows for happenstance interactions, we was interested to what extent these new acquaintances could potentially be expanding an author's access to occupational resources. If this is true, namely that the process of weblogging increases one's social capital, then we would expect the length of authorship to be correlated with an increased amount of online occupational access. Before we address this measure, there are a few caveats to the online/offline distinction.

First, a few subjects appropriately asked the question, "how am we supposed to know the professions of my online acquaintances?" This is related to the broader question of whether or not one can extract resources from online social ties without knowing specific details about their professional lives. The survey was meant to elicit the potential of an individual to extract resources, and understanding how large groups of anonymous individuals can pool social capital should be the subject of another survey unto itself.

Also, the definition of "know" for online ties was flawed. The survey defined knowing as, "if you saw this person on the street somewhere you could remember their name and start a conversation with them." Even if two people had met online, considered each other friends, and were aware of each other's respective occupations, there is a potential that they would not be able to recognize each other offline. The fact that "knowing" is defined by an offline interaction biases the subject's memory towards offline ties.

Finally, the question solicited a single answer for each occupation. To disambiguate the case that the subject knew multiple individuals in the same occupation, the survey specified that they should choose the individual they "communicated the most with," as a measure of tie strength. We chose this language because we knew that our other definition of tie strength ("someone you feel especially close to") would bias the results towards offline interactions; communication would select for the individual they were most acquainted with currently.

	Investment	M_{prof}	M_{soc}
PG_{on}	.226	.084	.144
PG_{off}	-.001	.057	-.007

N = 26,360. $p < 0.001$ for all correlations.

Table 10: Online and offline Position Generator scores. Control variables are Age, Gender, Education and PG_{on}/PG_{off} when observing the opposite variable.

Given all of these caveats, the relationship between the online position generator sum PG_{on} and the offline sum PG_{off} are shown in Table 4.2.4. In early analyses, we looked at the total sum PG_{sum}, but this value seemed to only be correlated with increases in either the online or offline scores, so we have chosen to present them instead. For each measure, we have controlled for age and gender, and to account for the tradeoff between the two, we have controlled for PG_{off} when looking at PG_{on} and vice-versa. These variables are shown with respect to the aggregate measures derived in the weblog and communication sections.

The largest discrepancy comes from the amount of total investment that the subject puts into his or her weblog; those individuals that invest a lot of time and energy into the practice are associated with a higher number of online access to occupational resources. The same is *not* true for offline ties, as an increase in weblog investment has no effect on offline relationships whatsoever.

The other peculiarity of these two measures is the distinction between M_{prof} and M_{soc}; while the difference is not overwhelming, there is a discrepancy between the various social capital measures and these two types of motivation. Those who are blogging for professional reasons tend to have a slightly higher offline social capital while online social capital is correlated with both, and more so for those motivated by personal reasons.

Despite all of the caveats, we were not expecting to find these interdependencies at all. Without a longitudinal survey it is impossible to definitively say anything about this relationship; any number of other variables could be the source of both weblogging behavior and increased online social capital. However, the size of this correlation reinforces the need to study online social capital formation in a more rigorous form, and with respect to any number of communication technologies. It also remains to be seen whether or not these individuals can actually extract the resources implied within the position generator.

5 Conclusions

This paper has presented two studies into the behaviors and resulting social effects of weblog authorship. We first observed the global structure of the blog readership network by indexing those weblogs updated over a one month period. The extracted network consisted of over 300,000 nodes and 1.7 million edges; the distribution of these edges as represented by in-degree followed a power law, suggesting that a large percentage of the attention within the community was governed by a few select weblogs.

Probably the most important contribution to understanding this community was the observation of a strong relationship between investment in the weblog and payoff in terms of audience size and feedback. As we anticipated, these measures of investment were shown to correlate very strongly with measures of attention and audience size. These data show that the weblog community rewards the author who puts time into their work, and that the length of one's blogging history does not solely determine their future audience. In this case, it is the hard-workers who get richer, not the previously-rich.

In order to better understand the meaning of this readership network, we conducted a large-scale survey of weblog authors. We contacted a random sample of those blogs with public email addresses from our aggregated data and also opened the sample to the general audience of authors. Two forms of blogging emerged from the survey section on motivations, professional and social, with differing behaviors and motivations; professional bloggers, in the minority, tended to have larger audiences and weaker social relations with their audiences. Social bloggers, on the other hand, had smaller audiences but more multiplex relationships. Furthermore, social bloggers were associated with the creation of social capital through online means while professional authors did not share this phenomenon.

Weblogs are a complex communication tool with a number of overlapping behaviors and social properties. The results of this paper start to separate these layers of relationships and present the community through distinct lenses. Through one perspective, we see professional writers trying to attract the largest audience possible, produce capital from their reputation and content. Through another we see a network of affiliated readership contingent upon real social ties. These online interactions mirror real, multiplex and often offline relationships, possibly leading to increased social capital. In order to strengthen our claims, we hope to replicate the results of our survey in a longitudinal fashion; after one years time we may be able to more closely observe the subtle changes happening in these authors lives.

References

Adamic, Lada, and Eytan Adar. 2003. Friends and neighbors on the web. *Social Networks* 25(3):211–230.

Adamic, Lada, and Natalie Glance. 2005. The political blogosphere and the 2004 u.s. election: Divided they blog. In *Workshop on link discovery: Issues, approaches and applications.*

Adar, Eytan, Lada Adamic, Li Zhang, and Rajan M. Lukose. 2004. Implicit structure and the dynamics of blogspace. In *Workshop on the weblogging ecosystem at the 13th international world wide web conference*. New York.

Bosnjak, M., and T.L. Tuten. 2003. Prepaid and promised incentives in web surveys: An experiment. *Social Science Computer Review* 21(2):208–217.

Ceglowski, Maciej. 2002. Blog census.
`http://blogcensus.net/`.

——. 2005. Languid: A language identification system.
`http://languid.cantbedone.org`.

Efimova, L., and A. de Moor. 2005. Beyond personal webpublishing: An exploratory study of conversational blogging practices. In *Proceedings of the 38th annual hawaii international conference on system sciences*.

Flake, G., S. Lawrence, and C. Lee Giles. 2000. Efficient identification of web communitites. In *Proceedings of the 6th ACM Conference on Knowledge Discovery and Data Mining*, 150–160. Boston, MA.

Van der Gaag, Martin, Tom A.B. Snijders, and Henk D. Flap. 2005. *Measurement of individual social capital*, chap. Position generator and their relationship to other social capital measures.

Ganzenboom, H. B. G., and D. J. Treiman. 2003. *Advances in cross-national-comparison. a european working book for demographic and socio-economic variables*, chap. Three internationall standardized measures for comparative research on occupational status, 159–193. Kluwer Academic Press.

Gibson, D., J. M. Kleinberg, and P. Raghavan. 1998. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*.

Global Reach. 2005. Global internet statistics (by language). Tech. Rep., Global Reach.
`http://www.glreach.com/globstats/`.

Gruhl, D., David Liben-Nowell, R. Guha, and A. Tomkins. 2004. Information diffusion through blogspace. In *Proceedings of the ACM Conference on the World Wide Web*. New York, NY.

Haythornthwaite, Caroline. 2000. Online personal networks: Size, composition and media use among distance learners. *New Media &amp; Society* 2(2):195–226.

Haythornthwaite, Caroline, and Barry Wellman. 1998. Work, friendship and media use for information exchange ina networked organization. *Journal of the American Society for Information Science* 49:1101–1114.

Herring, S. C., I. Kouper, J. C. Paolillo, L. A. Scheidt, M. Tyworth, and P. Welsch. 2005. Conversations in the blogosphere: An analysis "from the bottom-up". In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05)*. Los Almitos: IEEE Press.

Herring, S. C., L. A. Scheidt, S. Bonus, and E. Wright. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04)*. Los Alamitos: IEEE Press.

Kumar, Ravi, Prabhakar Raghavan, Jasmine Novak, and Andrew Tomkins. 2003. On the bursty evolution of blogspace. In *Proceedings of the ACM Conference on the World Wide Web*.

Lee, Su Hyun. 2004. Souped-up blog takes south korea by storm.
`http://www.iht.com/articles/2004/12/30/business/ptkorblog.html`.

Lin, Nan. 1999. Social networks and status attainment. *Annual Review of Sociology* 25:467–487.

——. 2001. *Social capital: A theory of social structure and action.* Structural analysis in the social sciences 19, Cambridge, UK: Cambridge University Press.

Lin, Nan, and M. Dumin. 1986. Access to occupations through social ties. *Social Networks* 8:365–385.

LiveJournal. 2005. Livejournal.com statistics.
`http://www.livejournal.com/stats.bml`.

Marlow, Cameron. 2003. Modeling emergent communitites through diffusion. In *Sunbelt International Social Networks Conference XXIII*. Cancun, Mexico.

——. 2004. Audience, structure and authority in the weblog community. In *54th Annual Conference of the International Communications Association*. New Orleans, LA.

——. 2005. The stuctural determinants of media contagion. Ph.D. thesis, Massachusetts Institute of Technology.

Marsden, Peter. 1984. Measuring tie strength. *Social Forces* 63:482–501.

Parks, Malcolm R., and Kory Floyd. 1996. Making friends in cyberspace. *Journal of Communications* 46(1):80–97.

Perseus Development. 2004. The blogging iceberg: Of 4.12 million hosted weblogs, most little seen and quickly abandoned. Tech. Rep., Perseus Development.

——. 2005. The blogging geyser: 31.6 million hosted blogs, growing to 53.4 million by year end. Tech. Rep., Perseus Development.

Rheingold, Howard. 1994. *The virtual community: Homesteading on the electronic frontier*. The MIT Press.

Smith, Marc. 1999. *Communities in cyberspace*, chap. Inivisible crowds in cyberspace: Mapping the social structure of the Usenet. Routledge.

Snijders, T.A.B. 1999. Prologue to the measurement of social capital. *La Revue Tocqueville* XX:27–44.

Wellman, Barry, and Milena Gulia. 1999. *Networks in the global village*, chap. Net surfers don't ride alone: Virtual communities as communitites. Boulder, CO: Westview.

Winstead, Jim. 2005. Blo.gs.
`http://blo.gs/`.

Wordpress. 2005. Wordpress weblog software.
`http://www.wordpress.org`.