Investment and attention in the weblog community

Cameron A. Marlow
MIT Media Laboratory
20 Ames St.
Cambridge, MA 02139
cameron@media.mit.edu

Abstract

While the weblog medium has grown out of a few modest technological innovations, the social and behavioral aspects of this emerging practice represent a large shift towards a new form of interaction: a massively distributed but completely connected conversation covering every imaginable topic of interest. This paper seeks to understand the social implications of hypertext links within the community, both from a systemic view of the entire network, and also from the perspective of individual authors. Using a large corpus of weblogs collected over a one-month period, an analysis of the structural properties of the weblog readership network is presented. These findings are compared to the results of an online survey of weblog authors around the topics of weblog use and linking behavior. Both cases suggest that the largest contributing factor to weblog audience size is the amount that an author invests into the practice.

Introduction

The medium of weblogging differs very little from other forms of online publishing that have been the basis of the web since its conception. A few years ago, only a handful of authors were writing daily to websites identified as weblogs, but undoubtedly there were many thousands of others who updated their personal homepages nearly as frequently, and in a similar writing 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.

Every informal social system has its own order, constituted by the attribution of friendship, trust, and admiration between members. These various forms of social association give rise to higher-level organization, wherein individuals take on informal roles, such as opinion leader, gatekeeper or maven. Among weblogs, much of this social order is observable through the hypertext links that authors make to each other's sites. These links come in two discernible forms: *static links*, typically made on the edges of a site, either as a sign of readership, support, interest, or marker of a social relationship and *dynamic links* made when an author links to specific content on another weblog, signifying a conversation or acknowledgment of their 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 content produced by the community.

Weblog links have been the source of a number of academic studies, with a number of hypotheses arising around the origins and dynamic qualities of the community structure. Some have suggested that these links are formed in the process of local conversations (Kumar *et al.* 2003; Herring *et al.* 2005), while others see them also arising from information diffusion (Marlow 2003; Adar *et al.* 2004). The distribution of readership across the community is not even, but rather concentrated on a few individuals, with most blogs only having a few links. How does authority and opinion leadership arise in this online community? Also, webloggers do not tell us what these readership links actually mean: are these people friends, acquaintances, or simply people they read? Do they communicate frequently in other media, or is blogging their only form of interaction?

This paper addresses these questions through an analysis of social behavior among bloggers in two parts. First, I will present the results of an analysis of the readership network extracted from a large corpus of weblogs collected over a one month period. The observed structure is shown to follow a power law, with the source of this distribution related to the frequency with which authors update their sites. Second, to validate these observations, I have conducted a broad survey of weblog authors, asking them a range of questions about weblog usage and behavior. The results from this investigation show that most of these links are not social, but more a sign of interest and readership. These measures of investment and attention are compared, showing again that the the more a blogger participates, the more attention they receive from other members. Furthermore, most of these links are shown to not be social at all, while increasing acquaintanceship is related to more regular communication using more media.

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. This research is too broad to be covered entirely here; instead I will present research dealing with models and analyses of link topologies and social structure. To understand the significance of weblog ties in the larger context of online communities, I will first review some applicable work in the areas of computer mediated communication (CMC) and internet sociology.

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 were coming 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, revolving around one or a few interests (Wellman & Gulia 1999). However, online relationships do not stay online forever; with continued social interest, users tend to increase the multiplexity (the number of simultaneous communication media) of their communication, and eventually meet face-to-face (Parks & Floyd 1996).

While much of the CMC research focuses on surveys or ethnographic context, 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; 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 & Adar 2003; Gibson, Kleinberg, & Raghavan 1998; Flake, Lawrence, & 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.

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 *et al.* 2004; Herring *et al.* 2004; 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 link similarity, content overlap, and timing of updates (Adar *et al.* 2004; Gruhl *et al.* 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 "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 the Blo.gs ping service, four random weblogs were selected, and from those weblogs an ego-network of alters 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.

A controversial essay published by Clay Shirky claims that the distribution of popularity and attention was disproportionately skewed towards a few individuals (Shirky 2003). He observed that weblog links follow a power law, and furthermore that this distribution was probably caused by the generative process known as *preferential attachment*, popularized by network researcher Albert-László Barabási. In his model, Barabási suggests that many power law distributions arise from the more connected nodes collecting more links, allowing for a feedback loop wherein the "rich get richer" (Barabási 2002). This inherent inequality gave rise to a heated debate as it suggests that popularity and authority are not caused by the quality or knowledge of the author, but rather by the amount of time they have been active within the community.

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 broad 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.

Design and Methodology

This section describes the methodology used in this paper. The first section speaks at length to considerations around the sampling of weblogs, followed by a description of data acquisition, analysis and survey techniques.

[1] Herring *et al.* (2005) have looked at this more closely, but their sample consisted of a qualitative sample of 24 weblogs.

Sample

Since there is no global system for tracking the existence of weblogs, there are a number of mitigating factors that help decide which frame population to use. There are four tried methods for acquiring weblogs: online *directories* where authors self-categorize their weblogs into a catalog, open *spidering* of the web using heuristics or statistical measures to identify blogs, large *applications* such as Blogger, Livejournal or Xanga where the population is specified, and finally *ping servers* where weblogs can identify themselves when they are updated. While most industry surveys have used applications (Perseus Development 2004; 2005), the most important features for this study are accuracy and recency. For this reason I have chosen to use ping servers to acquire the sample, with the most obvious choice being the popular public server, Blo.gs (Winstead 2005).

Aggregator

The first part of this study is an analysis of the link structures inherent in the content of weblogs. In order to produce a data set for analysis, I must first collect a large corpus of weblog content; this task will be executed by a crawling and indexing system built for this study. This tool will continuously monitor updates to the ping server and collects data about updated weblogs over time. The content of updated weblogs will be 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 referencing specific, deep content on another blog, or *static*, pointing to the to the front page[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, I 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 I have obtained from Blo.gs (Ceglowski 2005).

The aggregator was set to 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. Subsampling this network would remove the overall context, and looking only at individual nodes would take forever, but we can get a similar feel simply by looking at various calculable measures.

After the data set has been cleaned of all 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 (Cormen 2001) 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 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. There are a number of features that scaling might be contingent upon: frequency of posting, quality of posting, connections outside the network, and any number of demographic variables. In this part of the analysis I have looked to see if there is any relation between other variables I have observed.

Survey

The purpose of the weblog survey is to validate the observations made in the previous section, while adding context to the social implications of readership links. The survey will take 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* including all of those subjects who found the survey through public channels. The survey contains five sections in total, two of which are relevant to the topic of this paper: one on the meaning of links made between authors and a section on their use of weblog technology.

Weblog Use In the first relevant section, subjects were asked to detail their experience with weblogs along with the behaviors of their audience. To measure the level of commitment that a given subject had towards their site, the survey contained a number of questions about the time invested into various related activities. Weblog acts were 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 I wished to measure the effect that this input had on the popularity of their site, as quantified by their self-reported audience size and the comments received in an average week.

Links After subjects submit the address of their weblog, the content is fetched at that time and links extracted, in the same method used by the weblog aggregator. A set of 5 links are randomly selected. For each of these links, subjects were asked to classify the link into a number of different social categories (weblog, weblog post, personal homepage) or "other" for any other link. Subjects were then given subsequent questions about the link based on the type specified.

When authors specified that the link was made to another *person's* web page, the questions that followed were about the relationship between the subject and the weblog they linked to. The first question asks the type of relationship the

[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 & Tuten 2003), these figures imply a sample of 5,330 subjects or a round 5,000 with a response rate of over 21.3%

author had with the alter: friend[4], family, acquaintance, or "don't know them personally." The subject was 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.

Results

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.

Aggregator

The aggregator started collecting data on May 16th, 2005. Sometime during the month of May, Blo.gs was sold to Yahoo! Inc., unbeknownst to anyone in the weblog community, and despite my coordinated preparation with Yahoo!, the service was still unavailable between June 14th and June 16th.

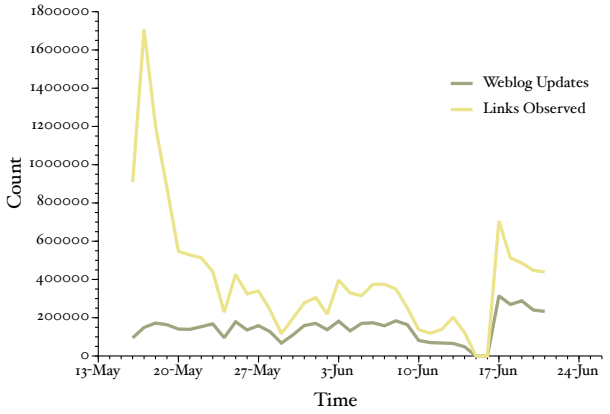

Figure 1: Weblog updates during the sample period

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 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 dropout of Blo.gs.

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

Language Of the 1,034,498 weblogs identified, 386,942 weblogs 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%. The list includes some surprising activity in a few languages. 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.

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, I first needed to clean the data to diminish their impact as much as possible through a number of steps of refinement.

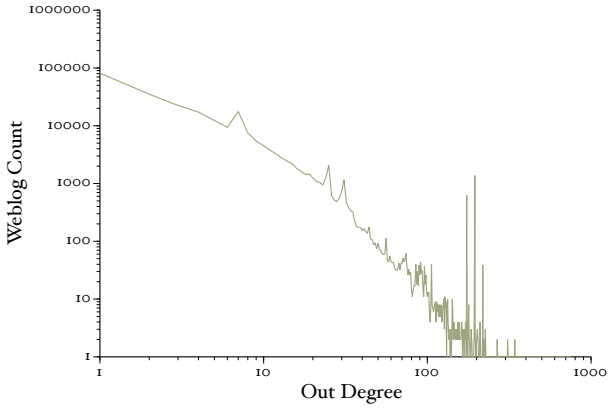

Figure 2: The initial observed cumulative out-degree distribution for the readership network.

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. I have deleted these weblogs, which consumed a full 85 of the top

100 updaters.

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

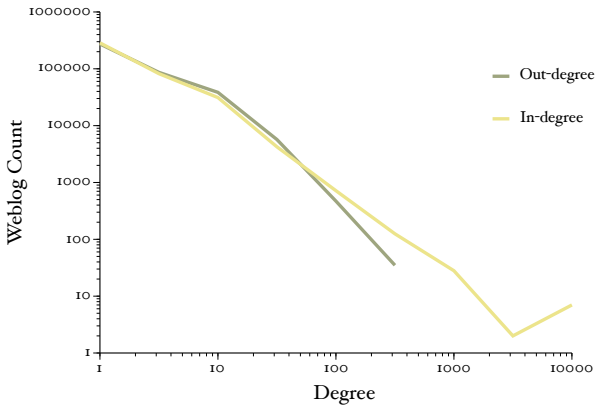

Figure 3: The initial observed cumulative degree distribution for the readership network.

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, excepting those that did not include the entire pre-configured set.

Connectedness To arrive at the components in a digraph, the graph is first converted to its undirected form and then searched using breadth-first-search (Cormen 2001, p. 532). 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 spidering approach to weblog acquisition, or a global ping server, they will remain isolated.

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, making the ratio 2.6:1. Given the short time frame of the study, I had expected this ratio to be much higher, especially since, accounting for aggregation over time, my last look at similar data suggested something in the range of 10:1 (Marlow 2004). Pearson correlations between the static and dynamic in- and out-degrees are shown in Table 1.

	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

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 implies that for variations around the mean of the distribution (which will be a low in-degree), the relationship will be so strong it will overcome the exceptionally large variance.

Investment What property exactly is it that determines popularity, either from a dynamic or static perspective? My 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 lists 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. Without any gauge of quality, it is clear that there is a strong relationship between the frequency of posting and a weblog's in-degree, and more-so for dynamic links than static. Intuitively this contradicts the notion that these power laws exist because of preferential attachment; if your degree purely related to the time at which you joined the network, then why is there such a clear relationship to the amount of information provided by these top sites? The fact that recent weblogs such as Engadget and Gizmodo can have such dominance also throws a wrench into the Barabási/Shirky argument. These data would suggest that there is more than just one force determining the

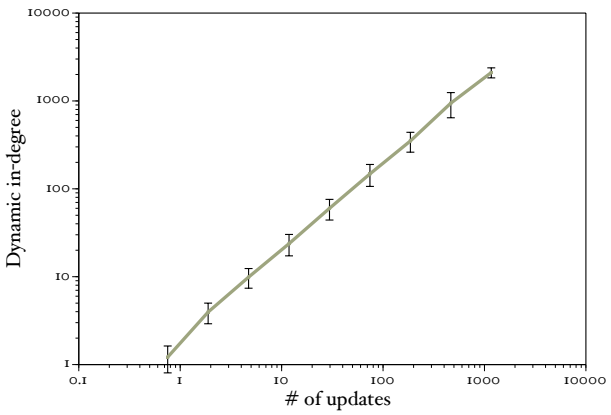

Figure 4: Updates vs. Dynamic in-degree

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. *Sex* is coded as 0/1 for male/female and *education* as 0 for "less than High School" with 6 as a graduate degree.

growth of these networks.

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.

A few caveats should be made explicit. First, the survey did not fully address issues of multiple authors. 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, I did not expect the response that I 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.

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 & Tuten 2003), a respectable response given the high probability for error in the extracted emails..

Table 2 contains the general demographic information (age, sex and education) for all three samples. Age was measured as the current year (2005) minus the year they entered as their birth year, sex was coded as 0 for male and 1 for female, and education was coded as 0 being less than High School to 6 being a graduate degree. 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.

Weblog use The first section addressed the various ways in which authors utilized their weblogs, and types of activities they engaged in on other sites. 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, as does the authors expectation of how long they will continue the activity.

	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.

Taking the sum of each investment measure we can each of these ordinal measures, I have constructed an aggregate value of investment. Figure 5 shows the relationship between this value and the self-reported audience size as a

scatterplot histogram. This plot complements the observations made from the aggregator data, showing a clear relationship between investment and audience size. Unfortunately, without longitudinal data, I cannot definitively determine 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.

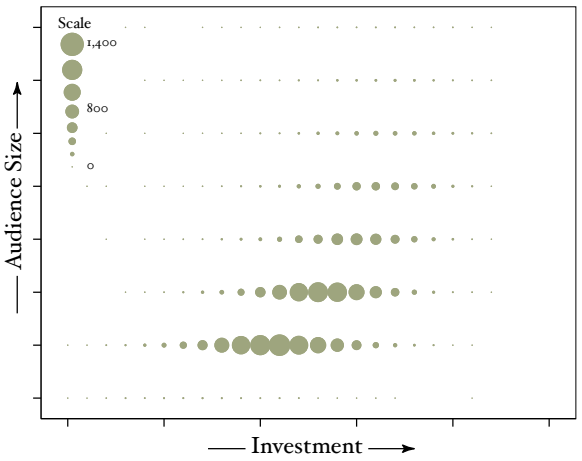

Figure 5: Relationship between author investment and self-reported audience size.

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

Of the 26,075 links listed as social links, exactly 9,700 came from a disregarded sample (LiveJournal or incomplete), and an addition 1,351 were missing data for all of the subsequent questions. After excluding these links, the data includes 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 issues: 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.

Table 4 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.

	Link Type		
Relationship	Post (%)	Weblog (%)	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 4: Social link type and relationship

	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 5: Readership and relationship

The next addressed is how many of these links are "live," denoting weblogs that the subject reads regularly and how many are "dead," pointing to readership that no longer exists. Table 5 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 I expected to find high readership for friends' weblogs, I was 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. As would be expected, 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 be some amount of bias associated with affirmation of the subjects self-image, namely that misremember their reading as much more frequent than in reality. they would rather remember having read these weblogs more recently than they may actually have. 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.

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.

Conclusions and Future Work

The data collected by the aggregator provided a rich network of relationships that could be used to understand the possible social structure within the community. 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. Both of the observed networks were almost entirely connected, having only a few components besides the largest one.

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 I 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 make stronger claims about these results, a longitudinal studies replicating the aggregator and survey are in order. This would allow us to evaluate the causality of a number of the observations made here.

References

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

Adar, E.; Adamic, L.; Zhang, L.; and Lukose, R. M. 2004. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem at the 13th International World Wide Web Conference*.

Barabási, A.-L. 2002. *Linked : the new science of networks*. Cambridge, Mass.: Perseus Pub.

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

Ceglowski, M. Languid: A language identification system [online]. 2005 [cited July 12, 2005]. Available from: `http://languid.cantbedone.org`.

Cormen, L. C. E. R. R. R. S. C. 2001. *Introduction to algorithms*. MIT Press, 2nd edition edition.

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

Gibson, D.; Kleinberg, J. M.; and Raghavan, P. 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). Technical report, Global Reach. Available from: `http://www.glreach.com/globstats/` [cited June 26, 2005].

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

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

Herring, S. C.; Scheidt, L. A.; Bonus, S.; and Wright, E. 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.

Herring, S. C.; Kouper, I.; Paolillo, J. C.; Scheidt, L. A.; Tyworth, M.; and Welsch, P. 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.

Kumar, R.; Raghavan, P.; Novak, J.; and Tomkins, A. 2003. On the bursty evolution of blogspace. In *Proceedings of the ACM Conference on the World Wide Web*.

Lee, S. H. Souped-up blog takes south korea by storm [online]. December 2004. Available from: `http://www.iht.com/articles/2004/12/30/business/ptkorblog.html`.

LiveJournal. Livejournal.com statistics [online]. 2005 [cited July 2, 2005]. Available from: `http://www.livejournal.com/stats.bml`.

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

Marlow, C. 2004. Audience, structure and authority in the weblog community. In *54th Annual Conference of the International Communications Association*.

Marlow, C. 2005. *The stuctural determinants of media contagion*. Ph.D. Dissertation, Massachusetts Institute of Technology.

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

Parks, M. R., and Floyd, K. 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. Technical report, Perseus Development.

Perseus Development. 2005. The blogging geyser: 31.6 million hosted blogs, growing to 53.4 million by year end. Technical report, Perseus Development.

Rheingold, H. 1994. *The Virtual Community: Homesteading on the Electronic Frontier*. The MIT Press.

Shirky, C. Power laws, weblogs, and inequality [online]. February 2003. Available from: `http://shirky.com/writings/powerlaw_weblog.html`.

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

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

Winstead, J. Blo.gs [online]. 2005. Available from: `http://blo.gs/`.

Wordpress. Wordpress weblog software [online]. 2005 [cited June 25, 2005]. Available from: `http://www.wordpress.org`.