Project number: IQP-CEW-0801

UNDERSTANDING PARTICIPATION IN WEB 2.0 SITES

An Interactive Qualifying Project Report

submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

by

Philip N. Hanson

Date:

Professor Craig E. Wills, Advisor

Abstract

The growth of dynamic Internet-based technologies and “Web 2.0” has placed web-based marketing in uncharted territory. We examine the efficiency of common Web 2.0 advice related to blogging and compare current weblog characteristics with those of physical meetings in order to identify methods of encouraging reader participation.

Table of Contents

1. Introduction

For nearly two decades, commerce over the Internet has been growing in volume and therefore in importance. Society as a whole has embraced the connected life, bringing the library, television, and marketplace on-line. Businesses new and old must work in a digital marketplace and, just as in the physical marketplace, appearance and interaction with customers can have a large impact on sales. In addition to products, ideas can prosper in the digital space when they are presented well. Traditionally, the process of presenting products or ideas in an appealing way is called marketing.

One might imagine that a decade of experience would perfect our knowledge of Internet marketing, but the Web is far from constant. Technologies periodically rise up, change, and fall away, and in the process they alter the landscape. Now the field is changing again as new kinds web sites – often called “Web 2.0” sites – alter the flow and sources of information on-line. In the process, these sites are creating new avenues for marketing products and ideas. They are connecting society in new and unique ways, or possibly perfecting old ways. They are providing the means for companies and non-profit organizations to interact with people in a personal way over the Internet.

But if these new technologies are tools, then where are the user manuals? What, precisely, is a podcast or a social network? What is the best way to use a weblog or a wiki for marketing? These questions have remained largely unanswered. Those who do offer advice frequently do not have data to support their conclusions.

This project examines just those questions. By defining and surveying the various types of Web 2.0 sites, then proceeding to focus on one type of web site – weblogs – we can attempt to extract innate “best practices” from the resulting data. In the process, we can examine whether our collected data appear to support or contradict common advice about the use of weblogs.

The overall goal of this project is to expose good techniques for Web 2.0 interaction, specifically within weblogs. Obviously no amount of technique can counteract a sufficiently bad product or idea, and it cannot replace good content, but good ideas and products can only benefit from well-executed interactions. Whether in person or aided by twenty-first century technology, the way people and businesses interact with each other is as important as ever.

In the remainder of this report, we will examine and analyze several facets of these new interactions. We discuss background information for our work in Chapter 2. Descriptions of our research focus, metrics, and definition of successful blogs follows in Chapter 3, and a description of our research methods directly succeeds that in Chapter 4. In Chapters 5 and 6, we will detail the gathering and analysis of weblog participation data. Finally, we give conclusions, advice, and recommendations for future research in Chapter 7.

2. Background Information

Before discussing how to use Web 2.0, it would behoove us to define what it is. In this chapter, we define Web 2.0 for the purposes of this paper and proceed to discuss its constituent elements and their origins. As our study focuses primarily on weblogs, we then discuss common advice that is offered regarding user interaction on weblogs as well as analogous structures or meetings in social science which have been more thoroughly studied.

2.1. Web 2.0

There has been much debate in the past few years over the meaning of the term “Web 2.0”. Part of the confusion may be due to the rather vague idea which it is meant to convey: according to Tim O'Reilly, who coined the term in late 2005, a major tenet of Web 2.0 is “the web as a platform” [1]. This is not a useful guideline when making distinctions on the Internet, so O'Reilly and his team composed a list of Web 1.0 web sites and their Web 2.0 counterparts, which we have reproduced here in part as Table 1. Each row contains a Web 1.0 technology or web site and its Web 2.0 analog.

Table 1. Web 1.0 Technologies and Their Web 2.0 Counterparts

Web 1.0	**Web 2.0**
Ofoto	Flickr
Britannica Online	Wikipedia
Personal web sites	Blogging
Content Management Systems (CMS)	Wikis

In O'Reilly's definition, Web 1.0 sites are not simply those which came earliest in the Web's history. Rather, they are those sites which simply display information instead of interacting with users or allowing them to interact with each other. Participation is lacking Web 1.0.

The entries on the right side of Table 1 can also be visualized as a Web 2.0 "space" which is composed of various categories of web sites, as in Figure 1. Categories in the left half of the diagram may be termed "content-based", and categories in the right half may be termed "technology-based" as an indication of their respective innovations.

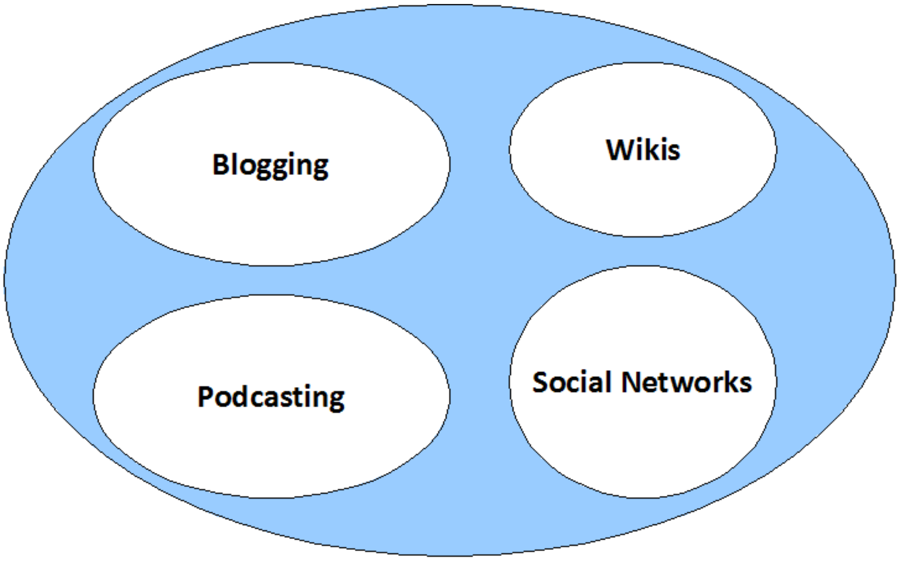

Figure 1. A Diagram of Web 2.0 Categories

As we will see, blogging and podcasting are not dissimilar from Web 1.0 content management systems and video downloads. In fact, the main point of separation between these "content-based" sites and their forerunners is the ability of users to respond in a public venue, keeping focus on the content. Wikis, social networks, and social bookmarking sites simply collect and organize content, leading us to call them "technology-based."

2.2. Social Networks and Social News Web Sites

While most of the other categories we discuss have predecessors in Web 1.0, social networking web sites hearken more to the days of electronic bulletin boards. As in bulletin board systems, the site operator provides the technology and hosting space for user content to be posted. User content may or may not be censored, but in general it is all viewable to the general public.

Notable examples are MySpace [2] and Facebook [3].

Similar in spirit are social news web sites, which allow users to submit news stories or web sites of interest. These links, along with selected metadata, are viewable by the general public. One well-known Web 1.0 analogue to this system is Slashdot [4]. One facet of Web 2.0 is community involvement, and social news sites clearly exemplify this trait [1]. Whereas Slashdot's content is managed by administrators – submitted articles never reach the front page without administrator approval – article ranking at sites such as Digg [5] is determined entirely by user feedback by an automated mechanism. Essentially, social networks and social news or bookmarking web sites are based more on technology than on original content. Nearly all the value provided by such web sites actually lies off-site, and little to no content is provided by the site owners or administrators.

2.3. Wikis

According to O'Reilly and his team, wikis descend from content management systems. Like their predecessors, wikis are a content-enabling technology. The main innovation of wikis is allowing users to manage all content in the wiki, rather than limiting access to administrators.

Wikis are technology- and user-centric in the extreme, granting full authorial control to all comers. Arguably the most famous example of this type is Wikipedia [6], which used to carry the motto “the free encyclopedia that anyone can edit.” While Wikipedia in particular is beginning to resort to greater editorial control on the part of administrators, it remains an excellent example of the wiki concept.

2.4. Blogging

Weblogs, commonly known as “blogs”, can be defined as “dynamic documents that contain periodic, reverse chronologically ordered posts on a common web page" [7]. They are similar in many ways to Usenet newsgroups: each blog is essentially a newsgroup focused on the perspective of a single person or group of people, and many – though not all – blogs allow viewers to leave comments on each post [8]. Like the contents of personal home pages before them, blog posts are generally not moderated.

Comments may be moderated, unmoderated, or disabled altogether, depending on the features of the blogging software in use and the preferences of an individual blogger [9]. Some examples of popular or influential blogs include Robert Scoble's weblog [10], Engadget [11], and Autoblog [12]. These three web sites are examples of an individual opinion blog, a technology-related blog written by multiple authors, and another multiple-authorship site on the subject of automobiles. The latter two sites are also categorized as "team blogs" due to their multiple authorship.

Blogs are distinctly Web 2.0 in form, as they make good on the mantra of "the web as a platform," but they are notably different from previous categories with regard to authorship. Whereas blogs have a single primary source of content, wikis and social networking or social news web sites allow all visitors to freely publish any content they desire.

2.5. Podcasting

Also called "webcasting", "audiocasting", and many other terms which end in "casting", podcasting is the distribution of a regularly scheduled audio or video recording by means of an Internet syndication technology such as RSS [13]. Such regular distribution was certainly not unheard of before the invention of automated Internet syndication, but prior to that time, it was necessary for users to check for new episodes frequently. Now programs such as iTunes and Juice can automatically download each episode as it is released and alert the user of its availability.

One could argue that podcasts are descended from blogs, for indeed they inherit nearly every feature except format. The RSS digital syndication technologies which power most podcasts were initially developed to deliver blog posts automatically [14]. Podcasts in general have a centralized primary content producer or team of producers. Often, podcasts have an accompanying weblog which alerts viewers to recently released episodes. In many ways, podcasts could be considered as audio or video format blogs, as opposed to the usual textual format.

Some podcasts have auxiliary web sites which allow for user feedback through blog comments, a forum, wiki, or similar web-based software solution. Such auxiliary sites complete the feedback cycle that is relatively integrated in blogging. The Revision3 network [15] and independent technology show Hak5 [16] are prime examples of this mechanism.

2.6. Best Practices for Blogging

Among the various kinds of Web 2.0 sites we have discussed, blogs stand out as the most conducive to studying author and reader interaction over the Internet. Therefore we focus primarily on blogs in this project.

As in all mediums, opinions have been advanced regarding the best methods for attracting readers to a weblog – and keeping them around. Few authors of such articles have advanced quantitative evidence of their opinions. The following list is a sample of blogging advice on the web:

- post articles on a consistent basis [17],
- make articles as short as possible [18],
- enable comments on your weblog [19],

- write long, informative post titles [20],
- keep sentences short and clear [20],
- include links to other weblogs and relevant web sites [18], and
- maintain focus in your content [21].

Some of the advice given is simply general writing advice, such as the above "keep sentences short and clear" and "maintain focus in your content." There is some weblog-specific advice, mainly regarding consistency and including references to other blogs. The first item above, posting on a consistent basis, is almost universally agreed upon.

2.7. Town Meetings

Because there is little existing data regarding user participation in weblogs and podcasts, it may prove useful to search out analogous situations. One can draw comparisons between blogs and shareholder meetings, for instance.

There is also an easy analogy between weblog comment threads and town meetings. This analogy is useful because there is existing research on participation within town meetings that may illuminate our examination of weblog participation.

Structurally, comment threads allow any registered visitor to participate in an ongoing discussion within and sometimes beyond the blog, and town meetings allow any registered voter to participate in an open discussion within the meeting. Blogs are simply more open, as some allow anonymous visitors to leave comments.

Given the strong similarities seen in structure, one might reasonably expect to see similar participation styles in blogs and town meetings. Frank M. Bryan and his team of students spent several decades gathering statistics on town meeting participation[22]. Some of this data can be analyzed to generate a cumulative distribution function which gives the percentage contribution – speaking at the town meeting, in this case – for any percentage of participants. Bryan's graphs are reproduced below as Figures 3 and 4. These graphs also include data for non-participating voting members which attended the meetings.

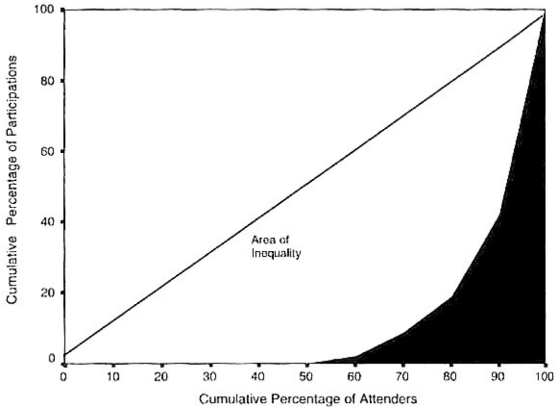

Figure 3. Contribution CDF for Starksboro Meeting 1997 (Plot 1 on page 156)

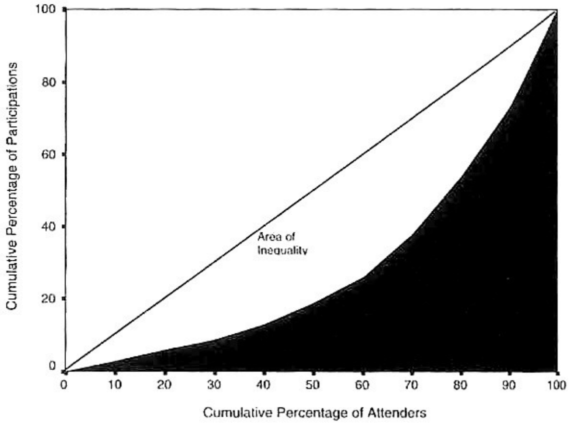

Figure 4. Contribution CDF for Panton Meeting 1982 (Plot 2 on page 156)

From these two graphs we can see that in the 1997 Starksboro town meeting, 50% of the participation came from 10% of the participants, while in the 1982 Panton meeting, 50% of participation was generated by 20% of the participants. While it is probably unreasonable to expect that a meeting of any size would achieve a perfectly equal ratio – half the participants producing half the conversation – comparisons with this ideal can be useful.

By calculating the closeness, or correlation, of the CDF graph with a straight line representing equality, we can obtain just such a numerical comparison. For the sake of comparison with later graphs, here we compute the correlation for that portion of each graph which accounts for participating attenders. Due to the nature of the graph, the correlation value must be between 0.5 and 1.0. A larger correlation value indicates closeness, so it is not surprising that the correlation for Figure 3 is 0.88 while that for Figure 4 is 0.93. From this we can say that discussion in the Panton meeting was more equally distributed among the participants.

2.8. Summary

We have seen in this chapter that Web 2.0, while vaguely defined in general, can be identified with several types of user-interactive web sites. Most Web 2.0 sites allow public user feedback and discussion of the site content. Some, such as wikis, are entirely user-generated. All of them are innovations on existing web technologies.

We also reviewed common advice offered regarding the management of blogs with regard to marketing and readership, finding that chief among them was a consistent schedule for new blog posts. In addition, we examined a study of town meetings, which are analogous to weblogs in structure. From this examination, we can infer that healthy on-line discussions might have a correlation value near or above 0.90.

3. Project Description

3.1. Research Focus

In this project, we focus on the use of Web 2.0 technologies by individuals, small businesses and non-profit organizations to communicate with and motivate people over the Internet. We exclude so-called Web 1.0 technologies and strategies from this report due to the large volume of existing research in that area, as well as the relative lack of research regarding this specific use of Web 2.0 technologies.

To guide us in determining effective strategies for using Web 2.0 to communicate with an audience, we examine the practices of successful web sites that compose the portion of the Web 2.0 space that is well-suited for the propagation of a particular opinion, message or messages via the Internet. These web sites draw content from both a centralized primary content producer and site visitors, usually in the context of a guided conversation. Sites which derive their value from technology, such as wikis, social networks, and social news or bookmarking services will not be examined in this project, because they rely only on user-generated content and do not depend on a centralized primary content producer.

Additionally, because there is no conceptual difference between technologies such as a blog without comments and a content management system, we may narrow our focus to include only those web sites that possess a feedback mechanism such as article comments or a web forum. Thus, our research focus lies squarely within those web sites that include an element of interaction between the primary content producer(s) and web site visitors, namely weblogs that allow comments and podcasts that have an associated feedback mechanism. This focus is illustrated by the highlighted areas in Figure 2.

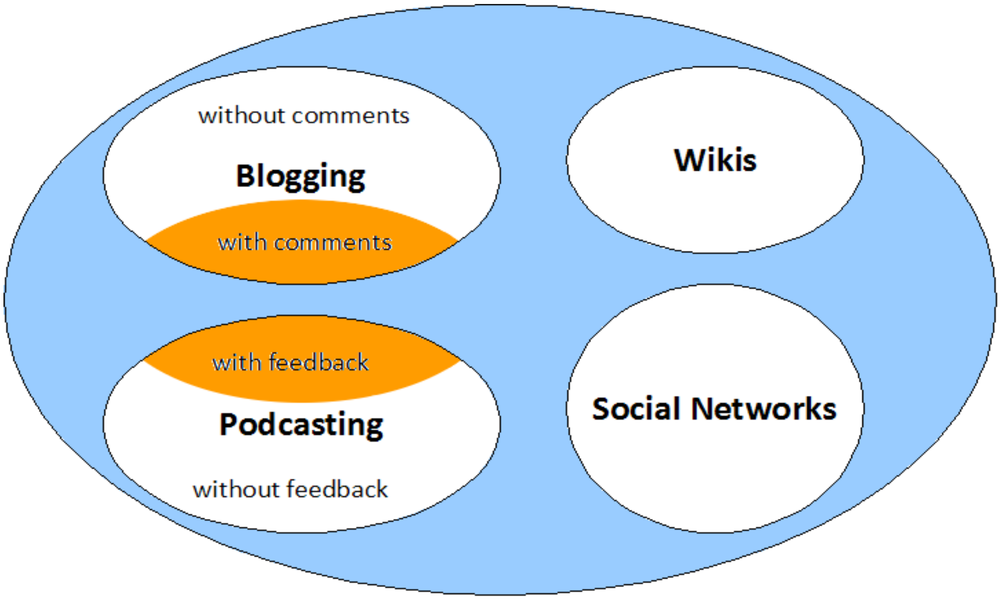

Figure 2. Research Focus

3.2. Metrics

Natural identification and measurement of success in communication is based largely on intangibles, such as “excitement” or “response.” Within our research focus, such ethereal metrics include “community involvement”, “popularity”, and “rate of return”, the latter being defined as the number or percentage of site visitors that take a certain action based on a request or recommendation from the primary content producer.

While the above metrics are largely unmeasurable by third parties, some can be approximated or measured indirectly. Many web sites, including some with a Web 2.0 focus, claim to measure the popularity of other sites. The popularity of a weblog might be measured by a combination of its Technorati[23] ranking and Google PageRank[24], for instance, or a podcast measured by its iTunes ranking and web site ranking together.

One metric peculiar to weblogs comes from the proliferation of RSS feeds: with the advent of comment syndication, it may be possible to generate community involvement statistics automatically. Measurements such as comments per post, per registered user, and per visitor may give insight into the character and quality of an on-line community. Applications of a similar nature have been constructed to monitor reputation and brand perception on the web [25]. There is also existing research regarding the characterization of weblog communities [26]. Specific cases of rate-of-return may be measurable using this technique, such as the rate of return on encouraging visitors to leave a comment.

More detailed metrics are likely to require site-owner cooperation. Rate of return on a book referral, for instance, would require cooperation from both the site owner and any number of on-line booksellers. Furthermore, such an approach will not be able to measure off-line purchases.

3.3. Success

Within the context of a conversation, whether on-line or in person, there may be many definitions of success. The appropriate definition depends on the goals of the speaker. Some conversations – and by extension, some blogs and podcasts – are intended to be secluded and private. In this instance, there are only a few contributors to the discussion, but each one accounts for a large percentage of the conversation's content. If this is the desired situation, success can be measured by the relative activity levels of each contributor.

Other conversations are large affairs, much like conference halls, where a few loud-mouthed participants generate the majority of responses, but a large and varied number of other contributors make a few comments each. Much like the excitement of a conference hall, excitement is generated here by the variety and sheer number of perspectives shared. In this instance, success is indicated by a large variety of contributors and not reliant on a vocal few.

Topic focus is also a consideration. Some sites and conversations are extremely focused on one subject area or interest, and these sites garner attention mainly from people who are interested in that subject. Others are totally unfocused and may develop a community based around discussion of the various topics that arise but not attract a people with a specific interest.

Because blogs are a public medium and most businesses and non-profit organizations desire to communicate with as many people as possible, we will consider a large or increasing audience size to be a partial measure of success for the purposes of this project. Additionally, topic focus may prove to be a viable measure of success if the desired level of focus is predictable.

3.4. Summary

By defining a research focus based on blogs with comments and podcasts which include feedback, we focus squarely on that part of the Web 2.0 landscape which is built on user interaction with a content producer. This is the situation in which most businesses and non-profit organizations which are concerned with user input will place themselves when starting a blog or podcast.

One difficulty here is that some useful metrics can be known only to the site owner. Without log files or direct database access, our knowledge must remain imperfect, and it is unlikely that many site owners would be willing to share such intimate details. This restricts the type of inferences we are to make to those which are discoverable through public sources. Also, we may not be able to say with complete certainty whether a particular site is successful or not, again due to incomplete information: we do not know the site owner's definition of success.

4. Methodology

Because there are so many opinions available about blogging techniques and little in the way of authoritative evidence, some method of generating an objective measure of weblog performance and characteristics is necessary. For this purpose, we wrote a set of utility programs to automatically collect metrics from an arbitrarily large number of weblogs.

Generation of blog data was split into two parts: data collection and data processing. During the collection stage, a custom utility downloaded the RSS feed for each blog being studied and placed the collected post and comment data into a simplified local XML file which was designed to be an intermediate file. The intermediate file's structure consisted of a single blog element which contained an unbounded number of post elements, each of which contained an unbounded number of comment elements. Post and comment elements contained attributes recording the item's author, publish date and time, and the length of the original data. The original item text from the RSS feed was not preserved.

Some feeds contain references within each post tag to the comments feed for that post, indicated by a Well-Formed Web (WFW) commentRss element[27]. When a feed contained these references, the collection utility also fetched the comments feed for each post. For completeness, only RSS feeds containing the WFW commentRss element within each post element or a supplementary comments-only feed were included in this study. After downloading the post and comment data, the collection utility merged new posts and comments with existing data inside the local intermediate file. Data collection was scheduled to be initiated automatically at noon and midnight.

After the collection phase produced an intermediate file for each blog under study, a set of custom data processing utilities could be used to generate statistics based on the file at any time. General statistics such as the number of post or comment authors, mean, median, and standard deviation of the time intervals between posts, post length, and number of comments could be calculated using a single utility and aggregated into a single file containing statistics for all blogs in the study.

A separate utility generated a cumulative distribution function (CDF) of commenter activity, i.e. the percentage of comments produced by a commenter and all more vocal commenters, arranged by individual. While generating the CDF, data were arranged with the most prolific commenters first. This ordering was an arbitrary choice, but ensured that graphs generated for the function would be useful for cross-blog comparison. The resulting function values were saved in a tab-separated-value file so as to be easily imported into a spreadsheet program for graphing.

5. Preliminary Data Run

5.1. Description

A preliminary run of the data collection process described in Chapter 4 was made during November of 2007 in order to test the collection process and tools. Data collected during this period was also analyzed in the manner of Section 5.2 below. This preliminary analysis helped identify ways to focus the later run and improve on the data-gathering technique.

Sites studied in the preliminary run had an RSS feed with WFW commentRss elements in each post, as the collection tool could not handle multiple feeds at the time.

5.2. Statistical Analysis

Statistics computed for this data set were the number of blog authors and commenters, number of times an author commented on their own post, length of posts and comments, time interval between posts or comments, and number of comments per post. Most of these statistics were condensed through averaging into a single value per statistic per blog. Additionally, traffic estimates from TrafficEstimate.com[28] were included. Table 2 gives a listing of sites studied along with relevant statistics.

Table 2. Sites Studied in Preliminary Data Run

Blog Name	Estimated Traffic [visits]	Num. Authors	Num. Commenters	Median Post Length [chars]	Median Comment Length [chars]	Mean Post Time Diff. [days]	Mean Comment Time Diff. [days]
TechCrunch	4613200	8	2048	349	171	0.04	0.04
XKCD	2319100	1	145	300	142	4.13	2.07
The Old New Thing	2072500	1	399	1760	336	1.00	0.14
ProBlogger	1862300	3	647	333	209	0.33	0.13
Smashing Magazine	1667300	1	549	1578	123	1.82	0.11
Larry Osterman's Weblog	1473200	1	137	2389	385	1.33	0.48
GigaOM	782900	7	305	339	250	0.08	0.09
CopyBlogger	756200	8	215	332	199	1.07	0.29
Think Progress	432500	6	498	375	232	0.03	0.02
Stuck In Customs	107000	1	62	275	74	1.17	1.06

The most natural type of analysis here is investigating the relationships between various statistics, also called correlation. Properly, correlation is a number calculated between two lists of values which ranges from -1.0 to 1.0 and indicates the degree to which the data points move in relation to each other, as well as the direction of movement. Thus, if two statistics have a large correlation value, any change in the value of one should induce a change in the value of the other. Only three pairs of statistics in this data set had a correlation greater than 0.4, which is a moderate value. Graphs depicting these relationships are reproduced here as Figures 5 through 7.

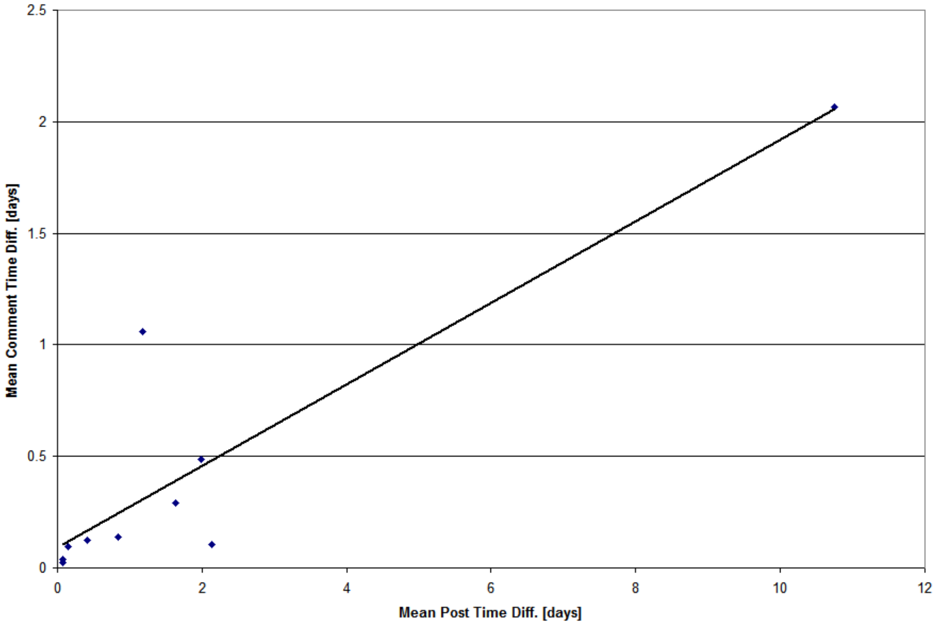

Figure 5. Mean Time Between Posts vs. Mean Time Between Comments. Correlation = 0.81

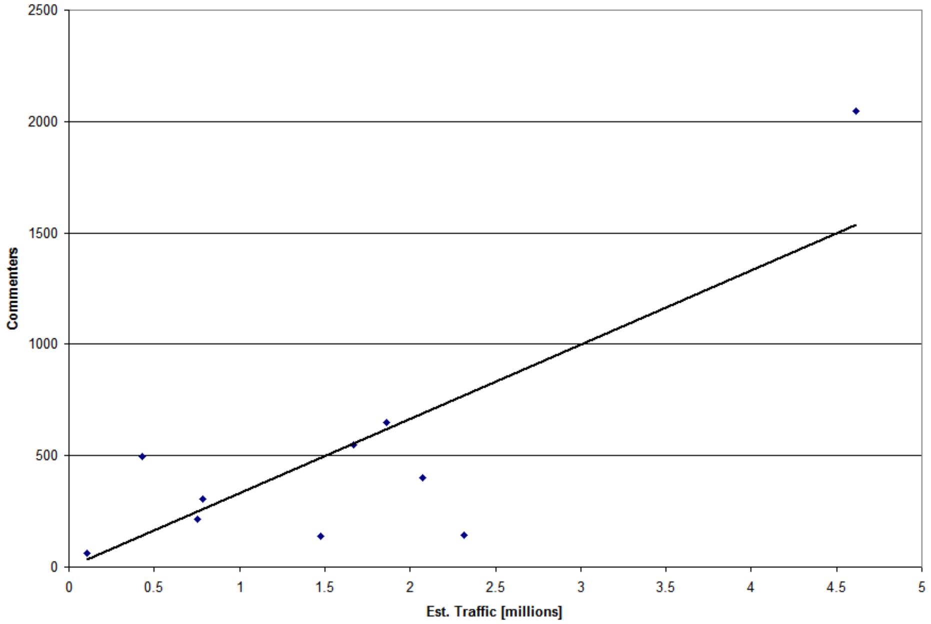

Figure 6. Estimated Traffic vs. Total Commenters. Correlation = 0.68

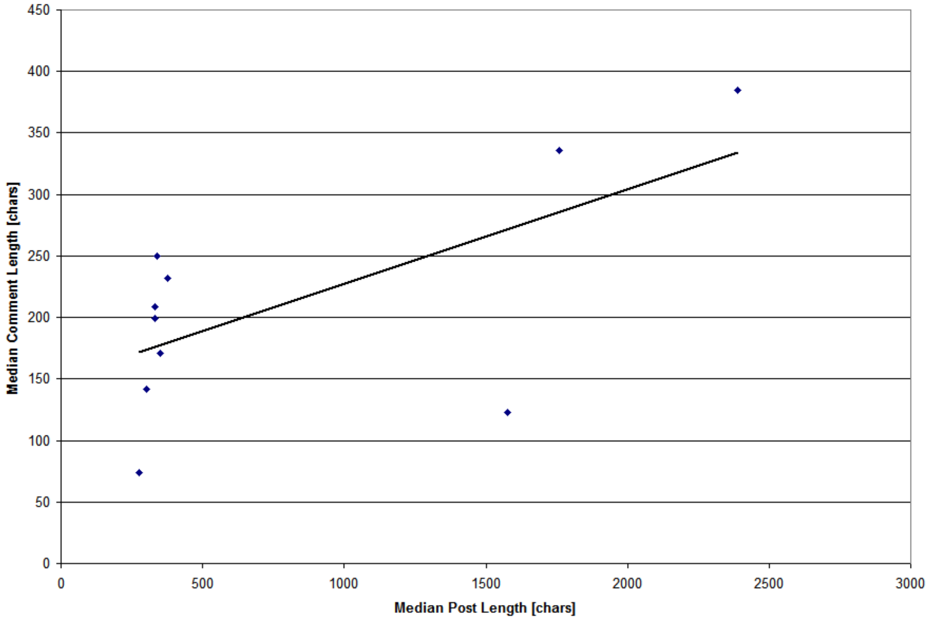

Figure 7. Median Comment Length vs. Median Post Length. Correlation = 0.41

We see from Figure 5 that there is a strong positive relationship between the average time between posts and the average time between comments. The correlation value is 0.81, the strongest of the set, and we can state this with 99% certainty within this data set. Thus, posting to a weblog more frequently inspired readers to comment more frequently within our set of studied sites.

Moving on to Figure 6, there is again a relatively strong positive correlation – this time with value 0.68 and certainty of 99% within the set. This indicates that the number of total commenters increases with increased traffic to the site, which is fairly logical and straightforward. It also suggests that there is some constant percentage of site visitors who leave at least one comment, regardless of site popularity. This percentage is likely to vary between sites.

Finally, Figure 7 relates the median length of comments to the median length of posts, for which there is a medium correlation of 0.41, stated with 96% certainty. It is possible that a multitude of short comments may have dragged the numbers down a bit, but given the clumping behavior of the graph, it seems more likely that there is a division within the data. Most sites in this data set published posts which were under 500 characters, while a subset averaged over 1500 characters per post. No sites occupied space in between 500 and 1500 characters per post. Further analysis with a larger data set could determine whether such a separation exists in general or only within this data set.

All other pairs of statistics yielded a correlation value less than 0.11, indicating a weak relationship, if any. We conclude that, for this data set, only three statistics can affect a blog's community: site traffic, length of posts, and frequency of publication. The other three statistics in the above relationships cannot be directly affected by a site owner or administrator.

Note that these correlations and confidence values are restricted to the set of preliminary data, because the volume of data within the set is not sufficient to be a significant indicator of the behavior of weblogs in general. Despite this, it may indicate the kind of results to expect in later data sets or subsequent studies.

5.3. Comment Density

Using the cumulative data from a blog's intermediate file, we can generate a Cumulative Distribution Function (CDF) which yields the percentage of comments generated by a given percentage of the total number of commenters. If every member of the community contributes equally to the conversation on a weblog, the CDF forms a straight line. In this paper, we call a straight-line CDF an “equitable” comment distribution.

Given a participation CDF, one may calculate the correlation between the function and a perfectly equitable distribution just as with any other set of paired data points. For the CDFs presented in this paper, the correlation value cannot fall below 0.5 and cannot rise above 1. Table

3 lists the correlation value and selected values of the comment CDF for all blogs in the preliminary data set.

Table 3. Cumulative Comment Distributions for Preliminary Data Set

Name	20% comments	50% comments	80% comments	Correlation
Smashing Magazine	11	44	78	0.997
XKCD	10	42	77	0.995
CopyBlogger	8	42	76	0.994
GigaOM	9	41	76	0.993
Stuck In Customs	5	24	67	0.963
ProBlogger	4	21	63	0.948
TechCrunch	2	17	63	0.939
Larry Osterman's Weblog	2	17	63	0.935
The Old New Thing	4	17	60	0.931
Think Progress	2	9	28	0.805

There are noticeable gaps in the correlation values which may be useful for categorization. We can say from Table 3 that blogs in the preliminary set fall into three main classes, which are delineated by correlation value as follows: Class I blogs have a correlation of 0.97 or above, Class II have correlations between 0.90 and 0.97, and Class III – one blog in this case – have correlations below 0.90. Examples from each class are shown in Figures 8, 9, and 10 below.

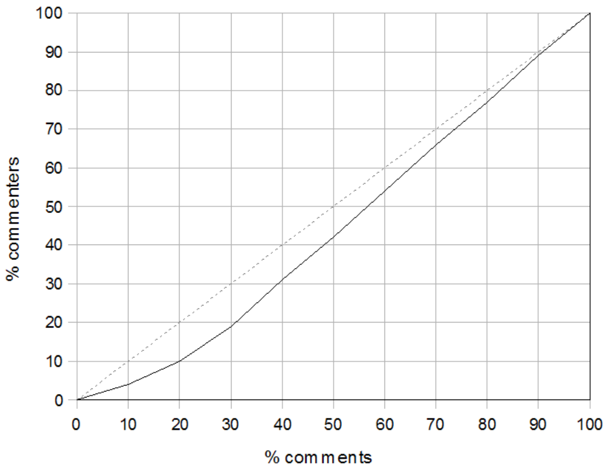

Figure 8. Participation CDF for XKCD.com

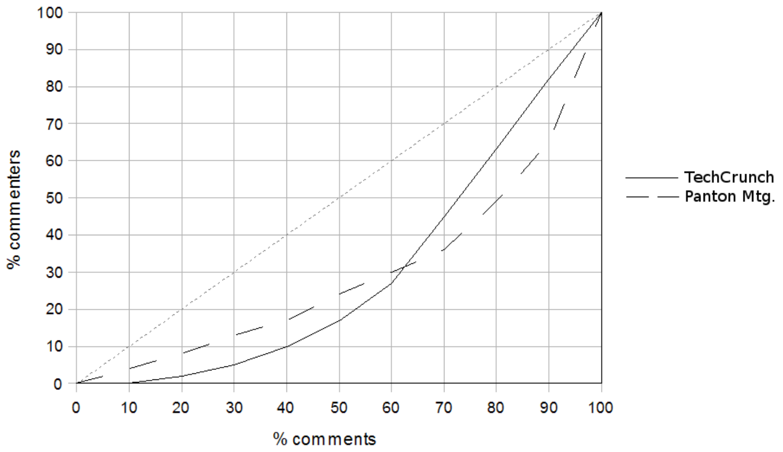

Figure 9. Participation CDF for TechCrunch.com with Data From Panton Meeting

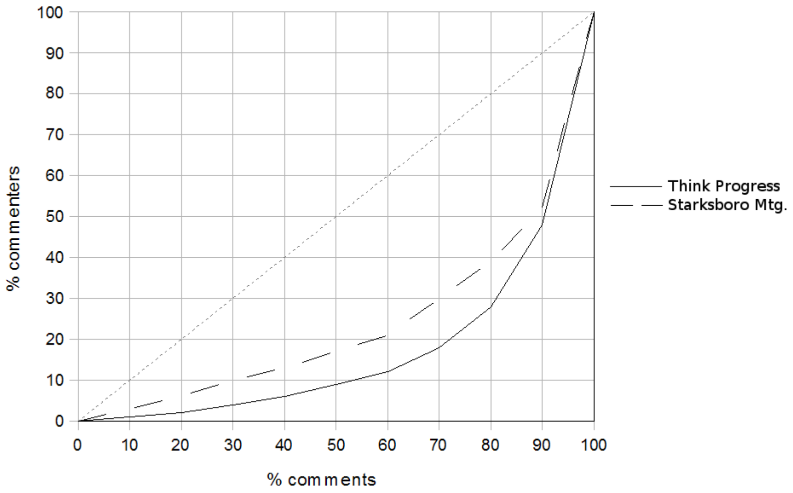

Figure 10. Participation CDF for ThinkProgress.com with Data From Starksboro Meeting

In each of these graphs, the light dashed line marks the level of perfect equality, the solid black line is the CDF in question, and the dark dashed line – if any – is a reproduced town meeting CDF from Section 2.7 for comparison. Just as when we calculated the correlation value, the town meeting data have been scaled to account for only those attenders which participated.

Figure 8 shows a typical Class I blog's CDF. The distribution is nearly equitable, and except for a short run-up near the beginning, it is practically a straight line. It is tempting to assume that this results from low traffic on a blog, in which case every user contributes a single post, but this is unlikely to be the case. While the example in Figure 8 is the XKCD.com “blag”, which is quite low-traffic in comparison to the popular web comic it accompanies, the same phenomenon can be observed in the comment threads of GigaOM, a popular and well-trafficked technology news site. XKCD's distribution has a correlation of 0.995 while GigaOM's distribution correlation is 0.993. Four blogs, or 40%, of the preliminary data set can be classified as Class I blogs.

Figure 9 shows the CDF of a Class II blog, TechCrunch, in comparison with the Panton town meeting CDF, which was included separately as Figure 4 in Section 2.7. TechCrunch's distribution has a correlation value of 0.939 and the Panton meeting's correlation value is 0.928, making them comparable in terms of participation. Bryan implies that the Panton meeting had the most equitable distribution of all meetings studied, so it is possible that blogs encourage participants to engage in well-rounded discussions more than town meetings do. Five blogs (50%) from the preliminary data set can be classified as Class II blogs.

Only one blog, or 10% of the set, in the preliminary set can be classified as a Class III blog. Its CDF is shown in Figure 10 along with the Starksboro meeting CDF, which is also included as Figure 3. Think Progress, the blog in question, is the only blog for which its CDF's correlation value, 0.806, is lower than the Starksboro meeting's CDF correlation value of 0.873. Interestingly, Think Progress is a political blog. This would seem to imply that given the right conditions, blogs can generate either more or less equitable conversations than town meetings.

5.4. Summary

Analysis of the preliminary data statistics suggests that there exists a relationship between the frequencies of posts and comments and that the number of active participants – commenters – on a weblog increases more or less linearly with increasing traffic. Additionally, long blog posts and long comments may be related.

Analysis of the comment densities suggests that, in general, blogs possess a more equitable participation distribution that other forms of open discussion, specifically town meetings.

6. Primary Data Run

6.1. Description

After rewriting the collection and analysis tools for better performance, a second data run with an expanded list of weblogs was performed during February of 2008. Based on findings from the preliminary run, metrics for average sentence length and the standard deviations of post and comment time intervals were added to the process. Though the collection and analysis tools were revised, the collection methodology remained that of Chapter 4.

6.2. Statistical Analysis

In addition to the statistics listed for the preliminary data set in Section 4.2, statistics for the standard deviation in time intervals between posts and comments and average sentence length were computed for the primary data set. Table 4 gives a listing of sites studied along with relevant statistics.

Again, we can calculate the correlation values for various relationships between these statistics. Revisiting the strong correlations found in Section 4.2, we find that all but one of the strong relationships from the previous set have disappeared, as illustrated by Figures 11 through 13. Figure 14 depicts the relationship between the standard deviations of time intervals between posts and between comments.

Table 4. Sites Studied in Primary Data Run

Blog Name	Est. Traffic [visits]	Num. Authors	Num. Commenters	Median Post Length [chars]	Median Comment Length [chars]	Mean Post Time Diff. [days]	Mean Comment Time Diff. [days]	StDev Post Time Diff. [days]	StDev Comment Time Diff. [days]
TechCrunch	3938900	10	1050	341	178	13.04	70.34	21608	371535
LifeHacker	2634700	6	414	2101	268	1.38	6.27	1311	58704
Mashable	2295300	9	1522	842	153	0.00	1.76	69	28476
XKCD	2130900	1	396	328	215	1.46	7.68	123	1548
The Old New Thing	2072500	1	603	1866	307	0.00	2.04	51	34584
ProBlogger	1862300	3	1285	329	235	0.13	2.91	263	5947
Larry Osterman's Weblog	1473200	1	132	1196	282	2.67	8.29	5366	27018
Smashing Magazine	1394400	1	1410	1660	106	1.75	15.90	1913	429583
ShoeMoney	1277900	5	450	291	101	5.83	29.24	3866	177176
ValleyWag	1156600	9	361	1939	171	0.42	5.65	1081	70929
Gawker	1073300	16	177	2448	159	0.04	2.54	282	81970
The Consumerist	962400	4	593	2494	284	0.00	1.75	407	18141
Matt Cutts's Weblog	953700	1	275	330	259	2.29	5.23	1652	75432
Neatorama	619400	17	1434	318	103	0.04	2.29	286	32786
GigaOM	619200	17	106	332	232	0.38	2.32	410	63507
CopyBlogger	559000	12	298	332	192	0.46	8.01	580	149867
Zen Habits	558700	1	1323	164	232	0.46	2.78	417	27454
Crooks and Liars	541400	15	3667	350	303	1.75	3.86	295	36250
LifeHack	517100	18	731	9421	223	0.83	2.49	2765	74448
Think Progress	410400	9	852	373	217	0.04	2.05	147	37363
Scobleizer	393200	1	1668	325	209	0.00	2.41	101	73647
Dooce	359900	1	83	1753	276	0.79	6.02	431	47626
Jauhari	158700	2	24	512	124	14.29	44.32	26523	407803
The Wrong Advices	124800	1	63	322	106	0.04	2.89	84	28612
Jonathan Coulton's Weblog	123800	1	844	309	152	0.00	2.61	408	74240
Stuck In Customs	107000	1	101	276	86	11.04	49.55	4133	420689
At0mica	43100	1	50	312	102	1.21	4.58	327	60713

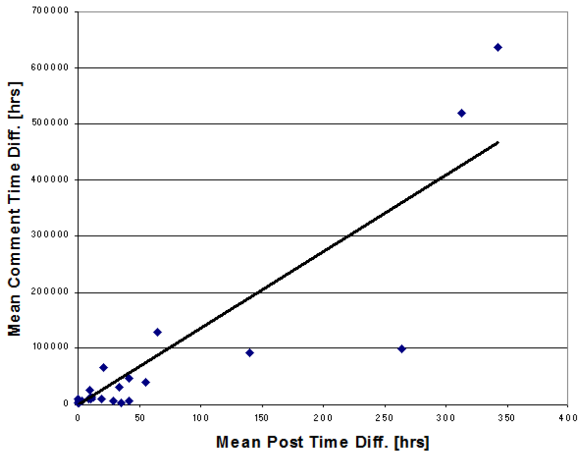

Figure 11. Mean Time Between Posts vs. Mean Time Between Comments. Correlation = 0.79

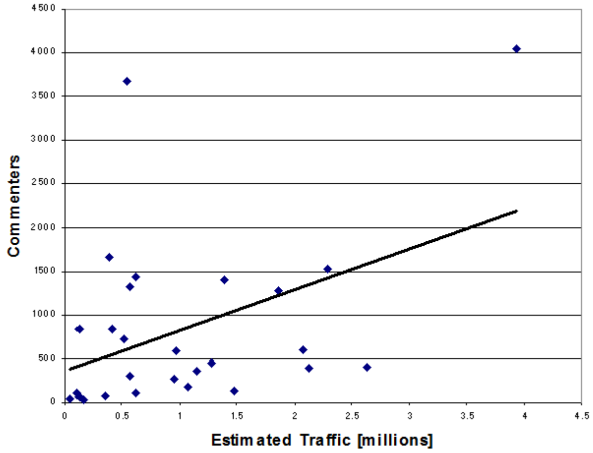

Figure 12. Estimated Traffic vs. Total Commenters. Correlation = 0.18

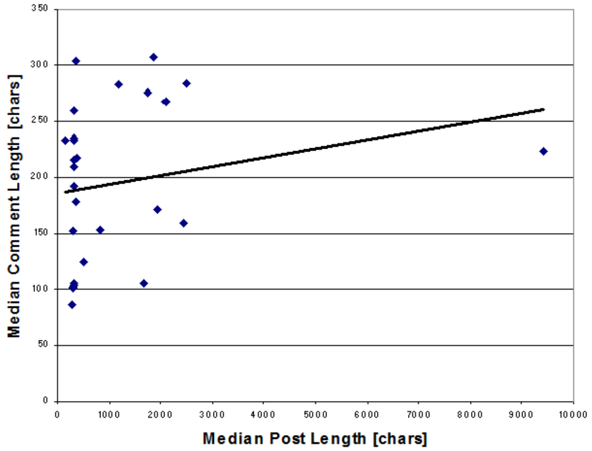

Figure 13. Median Comment Length vs. Median Post Length. Correlation = 0.04

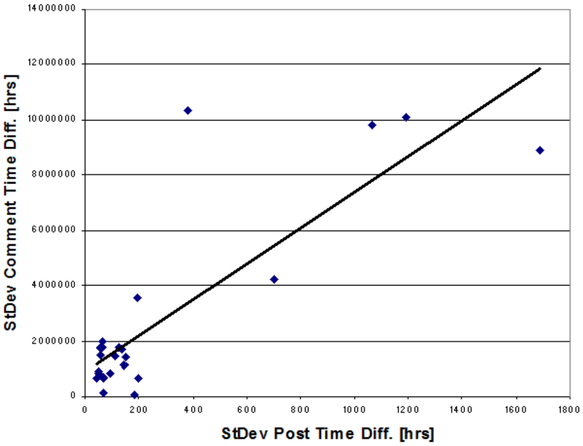

Figure 14. StDev Time Between Posts vs. StDev Time Between Comments. Correlation = 0.70

From these data, we see that the relationship between the mean time between posts and the mean time between comments, as shown in Figure 11, is largely unchanged: the new correlation value of 0.79 compares well with the previous one of 0.81.

The other two relationships, between estimated traffic and total commenters and between median comment length and median post length, shown in Figures 12 and 13 respectively, are no longer as strong. Traffic vs. commenters has a correlation of 0.18, down from 0.68 before, and the correlation for comment vs. post length is nearly nonexistent at 0.04 compared with the previous value of 0.41.

One pair of new statistics exhibited a relationship: the standard deviations of time intervals between posts and between comments, which can be considered measures of consistency, have a correlation value of 0.70. Together with the earlier referenced mean time interval statistics, this suggests that posting often and with a regular schedule inspires readers to comment often and regularly.

No other relationships exhibited a correlation value greater than 0.30. While it may appear that this data is less conclusive, we would do well to note that the previous data set was considerably more restricted in scope and thus more vulnerable to skewing. The primary data set is more statistically significant. Decreased correlation between statistics may indicate either that blogs in general do not change characteristics based on technique or that there is simply extra noise in the data set.

6.3. Comment Density

By processing the intermediate data files from the primary data run in the same manner as described in Section 4.3, we can again generate cumulative distribution functions for the comments on each weblog. Table 5 lists the correlation value and selected values of the comment CDF for all blogs in the primary data set.

If we again divide the results into three classes, the result is similar. Class I, with correlation values of 0.97 and above, accounts for eleven blogs, or 41% of the set. Class II has correlation values from 0.90 to 0.97 and claims another eleven blogs, which is another 41% of the set. The remaining 18%, or five blogs from the set, fall into Class III, which is defined by correlations below 0.90. Figures 14, 15, and 16 provide examples from Classes I, II, and III, respectively.

It is interesting to note that Crooks & Liars and Think Progress, both political blogs, rank far down in Class III and are the lowest in the set with regard to comment equity, yet Gawker and Valleywag, which are gossip blogs, remain firmly in Class II.

Table 5. Cumulative Comment Distributions for Primary Data Set

Name	20% comments	50% comments	80% comments	Correlation
Dooce	15	46	78	0.998
Jauhari	13	45	78	0.998
GigaOM	9	42	77	0.994
LifeHacker	8	38	75	0.989
LifeHack	7	37	75	0.988
Smashing Magazine	6	38	75	0.987
Wrong Advices	5	37	74	0.986
XKCD	5	35	75	0.984
CopyBlogger	5	34	74	0.982
Mashable	4	30	72	0.975
At0mica	5	28	71	0.974
Consumerist	5	24	68	0.963
Zen Habits	4	22	68	0.960
Gawker	4	20	62	0.945
Matt Cutts's Weblog	2	19	65	0.944
ProBlogger	3	19	62	0.939
Larry Osterman's Weblog	1	16	64	0.938
Scobleizer	2	18	58	0.937
Stuck In Customs	1	12	61	0.919
ValleyWag	2	15	54	0.914
The Old New Thing	3	15	55	0.914
TechCrunch	2	12	57	0.912
Jonathan Coulton's Weblog	2	11	50	0.895
ShoeMoney	2	9	53	0.886
Neatorama	1	7	44	0.868
Crooks & Liars	1	6	30	0.815
Think Progress	1	5	21	0.752

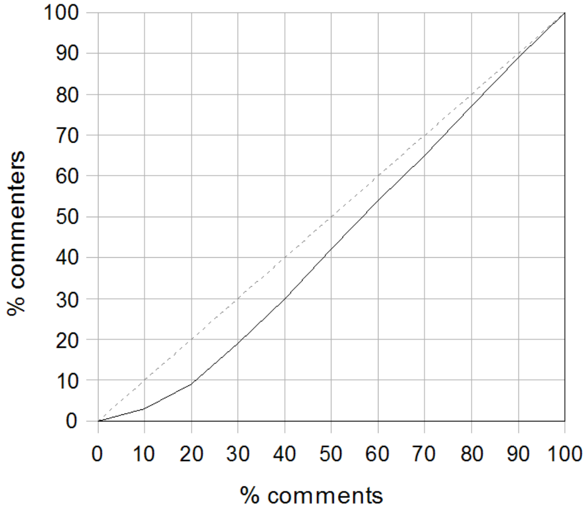

Figure 14. Participation CDF for GigaOM.com

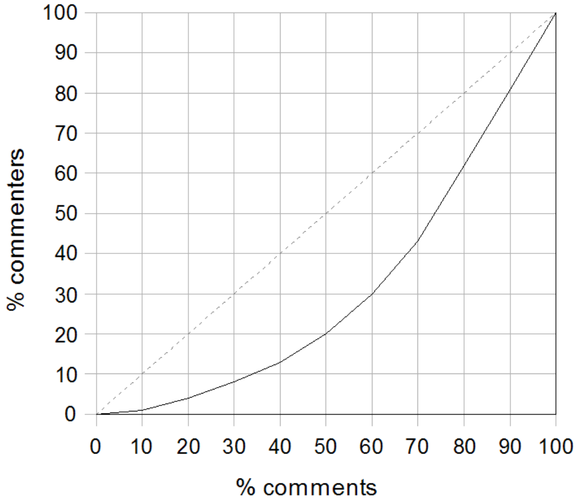

Figure 15. Participation CDF for Gawker.com

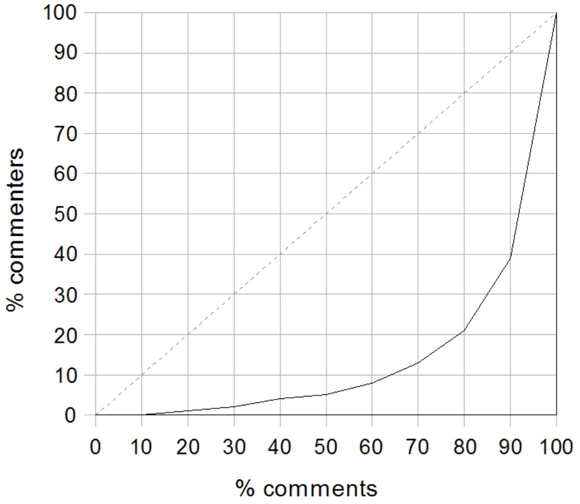

Figure 16. Participation CDF for ThinkProgress.org

From the data in Table 5, we see that comment distributions did not change significantly between data sets. This suggests that a blog's composition changes slowly. A majority of blogs fall into the same range of correlation values as town meetings, with some number around 40% venturing higher and surprisingly close to perfect equity among commenters.

Equity is again not tied to size, as the correlation between total commenters and comment correlation is less than 0.20. This indicates that, while more difficult to manage, it is certainly possible to develop large readership communities which engage in equitable conversation.

6.4. Summary

Reviewing these results in light of the blogging advice listed in Section 2.6 we see that, using the collected metrics, we can make conclusions about three of the seven items: the length of an article or of the sentences within an article has no measurable effect, but consistency in posting does affect reader participation. No suitable metrics were collected for two of the items listed, and the remaining two items are unmeasurable in nature.

The most oft-repeated advice for blogging, then, appears to be correct. Consistent publishing habits encourage reader involvement. Regardless of post length or writing style, the most important factor in building a community of readers around a weblog is consistency.

In addition, it is possible to develop a large but well-represented community of readers. None of our collected metrics indicate the method of development, and indeed it may be due to content or personality, which are not measurable.

7. Summary and Future Work

7.1. Conclusions

We began this inquiry into Web 2.0 Internet marketing techniques with a summary of previous methods and technology. Keeping in mind that not all phenomena can be explained by numerical analysis, let us examine the results of this project in light of their practical use. The data we collected indicate that the rate and regularity of comments can be affected by the rate and regularity of a weblog author's own posting habits. Thus, authors wishing to engage in discussion with their readers more often – one may presume more closely – should initiate the process by posting more frequently and consistently. This is in line with general advice proffered on the Internet, but it carries with it the weight of statistical analysis.

Some data also support the existence of a moderate to strong relationship between site traffic and the number of active commenters. The correlation is lesser when dealing with the highest-traffic web sites, suggesting that the relationship may be nonlinear if it does exist. Businesses can probably assume that increased traffic to their weblog will result in an increased number of commenters, but no assumptions regarding the amount thereof.

Readership composition is more of an open question, however. It is clear that a large minority of blogs have characteristics similar to those of town meetings and that a similar amount possess a more-or-less equal spread of participants. The remainder are heavily skewed by vocal commenters. What is not clear is which group is healthiest, or “best,” or whether such a distinction can be made. Equitable distributions appeal to our sensibility of fairness, yet heavily skewed distributions speak of passionate readers. Furthermore, the data did not indicate the cause of specific distributions nor how to reproduce them. It is reasonable to assume that the distribution is shaped by blog content, some set of hidden factors, or a combination of the two.

For individuals or organizations just beginning to blog, then, it appears that the best course is to post articles often and with a regular schedule while encouraging readers to leave comments. These practices will serve to create a baseline level of discussion from which later community efforts and reader interaction can take shape.

7.2. Future Work

There are many useful avenues of further exploration. Greater statistical significance and granularity of analysis can be obtained by using a larger set of sites with an expanded set of metrics. More accurate data could be gathered through collaboration with site owners by using a server-side tool to analyze the actual blog database rather than pulling data from sites through syndication.

A larger data set or different metrics might also indicate the optimal comment distribution for weblogs. If it can be shown that adjusting these metrics affects the distribution, this could fundamentally change the way weblogs are used.

References

1. O'Reilly, T., "What is Web 2.0". *http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html*, September 30, 2005
2. MySpace, *www.myspace.com*, September 2007
3. Facebook, *www.facebook.com*, September 2007
4. Slashdot, *www.slashdot.org*, September 2007
5. Digg, *www.digg.com*, September 2007
6. Wikipedia, *www.wikipedia.org*, September 2007
7. Chin, A., Chignell, M., "A Social Hypertext Model for Finding Community in Blogs". *Proceedings of the seventeenth conference on Hypertext and hypermedia '06*, August 2006.
8. Cohen, E., Krishnamurthy, B., "A Short Walk in the Blogistan". *Computer Networks: The International Journal of Computer and Telecommunications Networking*, Volume 50, Issue 5, April 2006
9. WordPress Documentation, *http://www.wordpress.org*, September 2007
10. Robert Scoble's Weblog, *http://www.scobleizer .com*, September 2007
11. Engadget, *http://www.engadget.com*, September 2007
12. Autoblog, *http://www.autoblog.com*, September 2007
13. Wolff, T. "Podcasting made simple". *Proceedings of the 34th annual ACM SIGUCCS conference on user services*, November 2006
14. Barsky, E., "Introducing Web 2.0: weblogs and podcasting for health librarians". *J Can Health Libr Assoc*, 2006
15. Revision3 podcasting network, *www.revision3.com*, September 2007
16. Hak5 podcast, *www.hak5.org*, September 2007
17. Alvin Huang's Web Marketing Strategies, *http://www.alvinhuang.com/bloggingblogging-consistently-for-traffic/*, December 2007
18. No Silence Here, *http://blogs.knoxnews.com/knx/silence/archives/2007/11/wapo_blogs_refl.shtml*, December 2007
19. About.com, *http://weblogs.about.com/od/bloggingbasics/p/Howtoblog.htm*, December 2007
20. Spiral Journal, *http://spiralbutterfly.blogspot.com/2007/11/ten-ways-to-write-great-blog-posts-that.html*, December 2007
21. WholesaleU Blog, *http://blog.wholesaleu.com/2007/11/29/top-10-things-to-avoid-in-a-weblog/*, December 2007
22. Bryan, F. *Real Democracy: The New England Town Meeting and How It Works*. University of Chicago Press, Chicago, IL, 2004.
23. Technorati, *http://www.technorati.com*, February 2008
24. Google Search Technology, *http://www.google.com/technology/*, February 2008
25. Ziegler, C., Skubacz, M., "Towards Automated Reputation and Brand Monitoring on the Web". *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006
26. Mei, Q., Liu, C., Su, H., Zhai, C., "A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs". *Proceedings of the 2006 ACM World Wide Web Conference*, 2006
27. Well-Formed Web Namespace Elements, *http://wellformedweb.org/news/wfw_namespace_elements*, January 2008
28. TrafficEstimate.com, *http://www.trafficestimate.com/*, February 2008