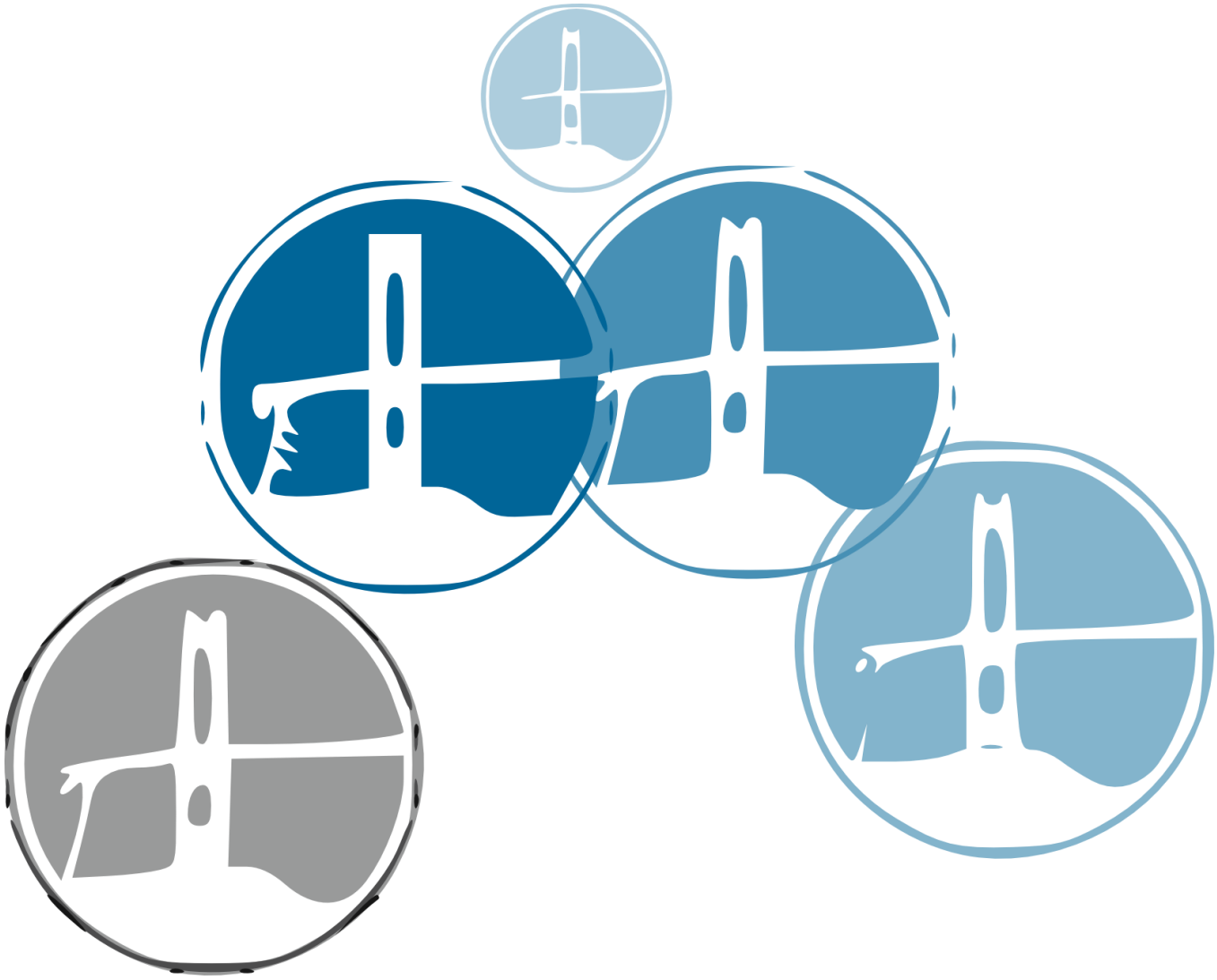

Durable Signal

Tracing a Freeform Community of Discourse on the Internet

DURABLE SIGNAL:

TRACING A FREEFORM COMMUNITY OF DISCOURSE ON THE INTERNET

Submitted to the

School of Interdisciplinary Studies

(Western College Program)

in partial fulfillment of

the requirements for the degree of

Bachelor of Philosophy

Interdisciplinary Studies

by

Thad Kerosky

Miami University

Oxford, Ohio

2007

APPROVED

Advisor _____

Nancy Nicholson

Abstract

This project examines the potential for a suite of methods which might be useful to assess the signal to noise ratio or more generally "quality" for an online community over time. In order to accomplish this, the whole archive of the community site Kuro5hin.org is downloaded and a statistical case study was performed. In studying the online community site Kuro5hin.org, the project has encountered the realities of managing a large and diverse dataset. As the project developed and the indicators were applied, I found interesting but tentative trends and thorny statistical questions. The project considered causality questions such as "what caused people to leave." Departure of contributors seemed tied to the change in quality. Active participation was indicated by boxplots generated to visualize the arrival, quartile and departure dates of prominent users. Some users who left the site early in its decline were surveyed. Academic sources provided content and conceptual architecture used to explain the decline of Kuro5hin discourse. Individuals this project consulted about Kuro5hin as a community shed light on the important aspects of communities online and why they fail. Choosing a suite of methods is a difficult problem that could require more analysis that will be pursued in the future. Ultimately, the project sketches the outlines of a larger proposal and gives a diagnosis of the Kuro5hin community.

Acknowledgements

For editing and focus I'd like to express my appreciation of: my parents, Nick and Joan Kerosky, Dr. William Newell, Dr. Christopher Wolfe, and Dr. Nancy Nicholson.

For pure ideas and brainstorming I'd like to thank Sean Duncan, Dr. Alton Sanders, John Fink, Dr. Leonard Mark, Dr. Sebastien Paquet, Dr. Alex Halavais, Derek Lackaff, and Dr. Donna McCollum. The group of K5 users I interviewed were profoundly insightful and quick. They improved my perspective a great deal.

Specific gratitude goes to Betty and Don Gerber for their wonderful digital equipment scholarship which gave me the computational horsepower to process the gigabytes of data required for this project at least four times as quickly as I would have otherwise, shrinking day-long processing jobs into afternoons. Nate Lamont Smith's extra laptop also lent significant technical support while I was editing.

General thanks goes to all my Western College community members who have supported me in my wide-ranging technical and nontechnical interests over these past four years. Any blame associated with the creation of these interests falls on the now scattered Kuro5hin community—everything from marmite to the singularity and through the philosophical implications behind "programming as art."

Rusty Foster, the founder of Kuro5hin.

Table of Contents

Table of Figures

Introduction

This project is concerned with the relationship between quality and the sometimes vicious dynamics of large gatherings of individuals online. Quality, in this context can be considered as the elegant appropriateness of text. It, for example has been crucially important to search engine algorithms since the beginning of the Internet. It is trivial for one person to write a search engine but almost hopeless for that one person to prioritize the results meaningfully. As difficult to assess as quality is, Google (and others) have demonstrated that its assessment can be a rich problem to solve. Quantitative quality has not been studied as it relates to online community. My thesis is that it is possible to develop a multipronged tool to evaluate the quality of discussion in an online community. These social gatherings occur naturally in the environment of the Internet—they result in community sites where people come together to exchange and discuss information. Wikipedia is one such community.

A wiki is the simplest database you can have at its heart[1]—a blank slate on the Internet that anyone who visits can modify. The wiki ideal and wiki software, of course, underlie Wikipedia, the Free online encyclopedia. Wikipedia is now the 10th largest website on the Internet (see Figure 1); it is a phenomenon which proves the strength and scalability of online community. Kuro5hin (often pronounced *Kero-shin*) and the democratic software framework[2] upon which it is based pre-dated the conception of the Wikipedia and are more structured than a wiki. Both the Wikipedia community and the

[1] Ward Cunningham www.wiki.org/wiki.cgi?WhatIsWiki

[2] The website software, Scoop was initially designed to allow voting on Kuro5hin stories but now supports hundreds of sites.

Kuro5hin community have a common goal of information sharing. There are two main differences between them which make the latter more interesting to study. First, the knowledge produced and contained by the site is much more compartmentalized thanks to the added structure. Second, the knowledge produced on Kuro5hin is linear, representing a single moment in the past seven years. Wikipedia spreads in all directions all the time and is constantly being revised.

The goal of this project at the outset was to collect a series of tests which together "triangulate" the elusive quantitative notion of quality to allow study of the thousands of community sites online. The project considers causality questions such as "what caused people to leave Kuro5hin" as they seem to be tied to the change in quality. For instance, during the donation drive on the site in 2002 the founder raised $30,000. This may have caused a change in the dynamics between the heretofore pro bono administration and the populace. Did this affect the way people related to the site? It may have encouraged users to become malicious and rebel against the founder and his newly exposed financial interests in their community. These rebels then may have encouraged others initiating the lengthy cycle of decline. It is quite difficult to tease out the accuracy of this one hypothesis of decline among hundreds. Though generally these kinds of speculation are outside the scope of the project, these are the sorts of questions that might be answered by this triangulation of tools.

An ethos is a structure of a specific consciousness (Geertz, 1993). It is the core that keeps visitors coming back to a community. Quality is part of this ethos and it is a moving target. Any site must be responsive to its users and its changing ethos.

The academic sources provided context and conceptual architecture used to explain the Kuro5hin discourse. The dynamic, ever-expanding world of the internet does not lend itself to thorough academic study. Most of the thinking done in this area has been pursued in a non-scholarly tone without significant academic consultation. Topics like social psychology and sociology loom over this area so this oversight was unfortunate. Individuals I consulted about Kuro5hin as a community shed light on the important aspects of communities online and why they fail. This project must humbly approach its corpus. Only a handful of tools are used and few causality questions like the one just posed are really answered. Searching for further meaning might make great strides given more analysis with higher powered methods in the same vein as those demonstrated. They should be pursued in future. As in the "Tragedy of the Commons" though, the solution is never only technical. Deep synthesis is required. In conclusion, I sketch the outlines of a larger proposal including more potentially valuable tools and give my diagnosis concerning the future of the Kuro5hin community given its past.

This project's title "durable signal" comes from the juxtaposition of two rich concepts. Durability is borrowed from Hannah Arendt's treatise on *The Human Condition*. Essentially **"durable" culture** is culture that is not packaged for mass consumption; rather it is something special that arises out of human interaction. At the heart of my analysis is the **signal to noise ratio** of the content. Signal consists of stories and comments that are valued by the populace such as relevant anecdotes or references that readers might wish to know about. Noise is content that muddies the

meeting ground making the signal scarcer and more difficult to locate. When this metric of interference shows disruption, a community has a good reason to disintegrate. Noise, as it relates to online community sites, might be divided into three classifications: poor semantics, tangentiality and hostile postings. The durable signal is the quality of the conversation within the boundaries of the community's ethos.

Quality is an elusive concept in any context, and the idea that the quality of discourse in an on-line community can be revealed by quantitative indicators may seem particularly dubious. Still I note that searching through data for a term is actually quite easy. The hard part of the searching problem is judging the quality of the items in the space at hand: Google shows that this quality call is not only possible but that it is worth getting right. I wish to use quality as a means to probe the timeline of gatherings of people online.

Object of Study

In this project, I intended to assess quality of discourse on the site Kuro5hin and whether it affects the rise and fall of the online community that developed around it. Kuro5hin (also called K5) is a community website where visitors can register for an account and then post an article on a live ballot, vote for or against other member's articles or comment on these articles. A scoring system allows users to put a value on articles. When an article reaches a certain threshold score it is posted to the site. Its contents are bounded primarily by its tagline: "technology and culture, from the trenches" (see Figure 8: Section popularity).

For the first four years of the site's existence this worked relatively well. A non-profit organization was formed to administer the site, not unlike the Wikimedia Foundation that manages the Wikipedia. The site's founder established a board of trustees. But after only their second meeting the community suddenly began to show deep signs of disintegration. It never recovered.

Appendix I, Figure 5: Example story exhibits a typical story posted to the site. The following two figures show rigorous comments from early in the site's existence (Figure 6: Ideal comments) and mediocre comments common in recent years (Figure 7: Average comments). Further figures show the constituency of the site in terms of user contributed content.

Ideally, the project would come up with a graph of quality over time for the Kuro5hin community. It might also be described in information theory terms as a signal-versus-noise graph. It is my assertion that this representation of the community is hinting to at least one major component of overall quality—the quality of durability. In this world of mass communication, durable communities are a fascinating and powerful phenomenon that brings together many of the best aspects of humanity. Internet communities are unique because a thorough backlog of information is available through site archives on intellectual and social activities. If a complete record has been maintained then the whole archive can be relatively easily downloaded for processing. It is then possible to walk back through the history of such a community and take notes on the quality of interaction.

Durable signal must be continually groomed and tweaked by administrators. It is tightly interwoven with ethos in that quality of the site experience is tied to process and content. If the site grows faster than the pool of administration available for this care then it is likely to fail.

Tools

An ideal tool for social researchers of the Internet age would be one where the quality could be charted using only the words of the archives and the time they were posted. This project is concerned with developing a prototype tool, as the ideal tool is not yet available. In 1997, Malhotra et al. observed that there existed a large body of literature on implementation of virtual communities but almost nothing written academically on the evolution of them (Malhotra, Gosain, & Hars, 1997). According to my searches, this observation largely holds even ten years later in 2007. In order to prepare to develop such a tool, I must take a preliminary step.

This step is a statistical description of Kuro5hin. The object of the quality analysis will be to identify possible causes and effects of the trends identified in the statistical description. In 1988, Williams, Rice and Rogers suggested that research methods for these new media could be an extension on traditional methods but proposed that researchers might find value in seeking out multiple alternative methods to try and triangulate their goal (S. Jones, 1999; Williams, Rice, & Rogers, 1988).

The triangulation tool takes in much data. For this particular community site there were 6,600 total stories and some 600,000 comments posted between December, 1999 and December, 2006. I downloaded these using a method called screen scraping.

Screen scraping is extracting the relevant blocks of text data from a website or other program while skimming off the irrelevant portions of the data. The screen scraping program examines every story and comment and copies and pastes the relevant text into the database which is then ripe for further examination.

In order to implement this prototype tool, various forms of data mining were employed. Data miners are interested in relationships within a corpus that are not immediately evident from any portion of the data. The favorite metaphor for managing and sifting through the overwhelming data is the spreadsheet. Though all of the data from the stories and comments were stored in a database they could not be easily analyzed in that form. Instead, large queries were executed to sum up the data into spreadsheets over periods like years, months, or weeks so that it could be analyzed using statistical software.

Text processing , or text mining is a subfield of general data mining. It typically involves turning long chunks of text into cells in the larger spreadsheet (Weiss, 2005).

Ultimately I envision a tool that would take the results from these data mining methods and combine them into a unified graph. I took several approaches to this corpus:

- **Activity:** Statistical analysis of the activity on the site over time, both on the macro, site-wide and micro, individual user scales (see Figure 16: Broad dataset figure on logarithmic scale);
- **Human:** Small scale, user interviews targeted by individual users' timelines (see Chapter 1: The Sky has Fallen);
- **Readability:** a simple readability/grade level measure of a text sample (see Chapter 5: Results and Appendix I);

- **Tangentiality:** a search for synonymous words indicating a topic common to both the story and the comments. I found that the next category was actually more precise in this goal and shifted my focus accordingly.
- **Semantic Tests:** Latent Semantic Analysis (LSA) has traditionally been used for automated grading of essays and semantic comparison of two texts (P. W. Foltz, Laham, & Landauer, 1999). It is more complex statistically and is framework-based (See Chapter 6 Discussion: Semantic Tests).

"Semantics" are only concerned with the meaning of a document, not the syntax and construction. Unfortunately, due to time constraints these LSA tests were only preliminary. I also take advantage of the anecdotal ratings made by visitors to the site. My graph of community health is incomplete but there are definite trends shared among these various methods identified in the discussion. Specifically, in mid-2003 there was a distinct crash in nearly all of the assessed curves (see Results and Appendix II).

Studying Online Communities

In the first chapter I provide a contextualized introduction to the specific community at hand. The goal of this chapter is to supplement the automated work of the later sections with definitions and brief qualitative analysis. I took the mined data concerning several users' date of departure from the site, chose certain interesting ranges, and asked for their opinions on the community's demise. These user perspectives proved valuable in my analysis.

Chapter 2 is an introduction to the academic conceptions around online community. It begins with a brief discussion of the various stages of academic literature

available on this topic. The rest of the chapter addresses the question of whether online communities can ever be as durable for participants as offline communities. Durability is defined in this chapter as it relates to the sorts of social disciplines consulted by this project . If participants find a community useful and engaging, it continues. Finally the ideas around and implications of Lurkers, or passive participants are discussed.

Chapter 3 is a brief chapter describing the elusive nature of discourse online and the assertions this project makes to ground it. I am interested in a signal hidden in much noise. Rigor must be addressed as it relates to casual intellectual conversation taking place online.

The Chapter 4, the methods chapter, first describes the mechanics of the different quantitative tests and how their results are important to my study. These are divided into four rough categories: readability, human structurally-inherent ratings, tangentiality and semantic tests. The chapter concludes with a section that goes into the explicit procedures accomplished including: the user growth curve gleaning process; the web server visitor and hit data acquisition; the downloading of the content; scraping content into the database format; the readability assessment and the latent semantic analysis exploratory tests.

Chapter 5, the results chapter, flows naturally from this analysis. First I describe the sketch of the community that the initial data scraping or data processing provided. Subsequently I discuss the statistical techniques I used, potential pitfalls in these analyses and very briefly, future analyses that could be performed.

Chapter 6 is devoted to general discussion on how and why communities fail. More specifically it is about how the site at hand disintegrated. Chapter 2 includes some speculation on that topic. This chapter attempts a wider synthesis of academic writing and contemporary community practices. It makes use of the results of the data and text mining techniques.

In the conclusion I consider the successes and failures of the model I proposed above. Future work on the model is a central point of these closing remarks. Finally I give my diagnosis of the Kuro5hin community. The durable signal was contaminated by the noise of hostile posting, tangentiality and poor semantics. These factors taken together with administrative inadequacy caused the decline.

The first appendix exhibits a high quality story and examples of positive and average comments on the site. The second appendix is rich with charts and graphs plotting the community out in a wide variety of figures. Even though some of these are not explicitly cited, I found all of these especially useful as I was thinking about the Kuro5hin lifecycle and constituency.

Chapter 1: Kuro5hin and Its Surroundings

There are many Internet sites which exist only for their users to submit content. A community site, in the context of this project, refers to a place online where users, typically of their own free choice, submit textual media. The community site hosts much value for its participants. Value is reflected in the qualitative fulfillment of certain goals of the community (Lackaff, 2005).This value is precarious. It is primarily based on the contributions of the posters and story writers. It might be expected that the more participants, the higher the qualitative value. In order to attract participants, though, there must be a common purpose of sharing ideas to begin with so this community site is to some extent a system feeding back into itself. All participants in a community site share a history with the site as long as they are active.

Common Structure

Before explaining the community at hand some general terms will be introduced. These terms enable any specific features and analysis to be discussed in the larger perspective of community and outside of any individual site.

A “user” or “poster” is an individual on the site who at some point has contributed information to the community. Primarily, this added information consists of comments or stories but contributing ratings on those comments or stories also require usership. On Kuro5hin and most other modern communities, users are granted pseudonyms. Occasionally, anonymous posting is also permitted, such as on

Slashdot.com. This aspect of community unique to Computer Mediated Communication (CMC) sometimes encourages people to separate their physical and online senses of self (Lackaff, 2005; Turkle, 1995). Individual participants will occasionally even have multiple users. This is discussed further in the following chapter.

A "story," as I will use it, is a post on the Internet which attracts users to critique the post and discuss issues related to it. I will use article as a synonym of story. Stories are often, but not strictly, longer than any single portion of the discussion surrounding them. In an ideal case, the story sets the tone of the discussion in the comments. A story should be contributed to by the people who regularly use the site. After contribution it is subject to certain publicly exposed value filter conventions set on a site-by-site basis. It could be checked and authorized by a group of moderators who review all of the submissions (Slashdot); voted on by a certain threshold of users (Kuro5hin) or be selected by a certain formula taking overall interest and momentum into account (Digg[3], Reddit, del.icio.us/popular).

An ideal "comment" or post is a contribution of additional information to the story. Since there are often many more commenters relative to the person or handful of people who have contributed to the story, the chances are reasonable that someone has expertise in the subject at hand and may add more value to the article than the original story provided. "Heterarchical moderation (whereby many, most, or all community members are given a small amount of power and responsibility for maintaining social norms and useful discussion)" is a common technique among

[3] The nature of these sites is explained in the "Cultural Aspects" and "Democracy" subsections.

community sites (Lackaff, 2005). Comments with a low rating are often off-topic for a variety of reasons (flamebait[4], troll[5] or no relevant information). There is always some degree of opinion expressed with ratings so unpopular comments may occasionally be "down-rated" purely on an ideological basis.

These comments may be further divided into threads for better organization of responses and especially responses to responses. It is reasonable to expect some level of cohesion between the "parent" comments and the "children" comments in a thread but this is outside the scope of the text analysis here.

Additionally, the text analysis portions of this project focus specifically on stories of significant length (typically between 500 and 1100 words[6]). Comments are expected to be shorter but are better analyzed when they have a length of 25 to 150 words.

Environs

The Kuro5hin community was founded on December 20, 1999. At the time of writing, this is a seven year legacy. Like the bustling communities of MOOs (MUD[7] Object Oriented) in the late eighties and early nineties such as LamdaMOO, MediaMOO, and the WELL or the multitude of Usenet Newsgroups, it did not rise and fall in a vacuum. The World Wide Web played a significant role in those disintegrations (Renninger & Shumar, 2002). After hypertext arose in the mid-nineties, the Internet became a much more complicated place. It is no longer so easy to point primarily at one change in the landscape to explain a crash in participation and quality. The websites

[4] Inflammatory post
[5] See Spam, Trolls and Crapflooding subsection.
[6] 50% of the stories are in this range, 25% fulfill and exceed it.
[7] Multi-User Dungeon, text-based chat rooms with environment.

with similar function might be expected to exemplify the larger trends, share and take users but also reflect the forefront of interest as frontier community-based sites burgeon.

The technological inspiration of the site was primarily Slashdot, a moderated online technology and science news site established in 1997[8]. It has been affectionately known as The Other Site within K5. By the time Kuro5hin was founded Slashdot already had thousands of users and had already published almost 10,000 bits of news[9]. Moderation for the posted stories was managed by a small team that approved their favorite submissions. Moderation for comments was achieved by granting a small portion of the site user base five points to spend on any given day in order to "mod-up" comments. If a given user has been highly rated many times he gets a high karma and automatically gets a free point attached to postings.

A site called Suck.com was a precursor to the K5 user-oriented format though may not have directly influenced it. It was a static HTML opinion column website started in 1995 which accepted user contributions. Plastic.com was spun out of that in 2001 and it took a fairly similar format to Kuro5hin with user-submitted stories.

Weblogs existed when Kuro5hin first started but the concept was still in its infancy. The idea is simple: a dynamic web page to to which a user can easily add content in blocks of text with the newest content at the top of the page and older content below it. In general a blog is a site all its own, sharing its visitors with no one[10].

---

[8] http://kuro5hin.org/?op=special;page=random
[9] http://en.wikipedia.org/wiki/Slashdot
[10] It is worthwhile to note that some of the most popular blogs have a closed *group* of authors.

By about 2002 blogs were becoming prevalent and by the time the site started to show signs of disintegration in mid-2003 they were everywhere.

Early on, as blogs were becoming popular, a special story section called Diaries was added to the Kuro5hin system. They did not require vote approval and could be easily published upon. They allowed for collaboration on stories and for personal accounts of events which might not be important enough for the main story section. Diaries encouraged collaboration better than weblogs in many cases since links to the diaries showed up on the main page. Since the diary is not completely controlled by a biased individual who is posting the stories, the commenter are more empowered to post what is on their mind and intelligently (or otherwise) criticize the author.

Spam, Trolls and Crapflooding

Spaming, trolling and "crapflooding" happen in most communities around the Internet. Different participants in any given online discussion have different tolerance for cognitive effort required to process the Computer Mediated Communication (CMC) (Q. Jones, Ravid, & Rafaeli, 2001; Lackaff, 2005). An unmoderated community will often be completely destroyed because the individuals involved are not willing to wade through the noise looking for the signal. Surprisingly spam is less of an issue on sites like Kuro5hin. Unlike email lists the Heterarchical rating systems on these community sites makes it trivial to recognize self-promotion and then rate it poorly. Users are good enough at this that such spam comments rarely stay for long. Editors also can to some extent ban users abusing the site in this way.

Trolling is tricky to define but it is generally posting solely to elicit a specific response from another user. Some of the users surveyed considered clever trolling to be the defining factor of intelligent and playful discourse on K5 while others wondered if the culture of trolling was what caused the downfall. A distinction must be drawn on the site between trolling and "crapflooding." Crapflooding is the destructive activity of posting many vacuous comments which are derogatory or completely off topic. They often contain jokes that only users embedded with other crapflooders or trolls would understand. The result is a clique of highly motivated but exclusive users maliciously wreaking havoc. This activity has been rampant since 2003 and is can be considered as the main detractor from discourse in the community since that time.

Trolling and "Crapflooding" are much more prevalent activities compared to spam. They present unique problems for public, Heterarchical rating systems. Malicious participants who are crapflooding are also the ones who are most interested in rating. Trolls often get rated up for being especially clever. The rating system can be overrun by these malicious or piqued users. Trolling might cause some trouble in my text analysis.

Cultural Aspects

Slashdot and Digg started with a strong user base of computer and Internet enthusiasts. Many of them come from an Information Technology background. Kuro5hin was different from these because it brought culture into the mix. Rusty Foster, the site creator, while deeply interested in technological issues and programming, pursued an academic career in the humanities.

Politics is a popular topic whenever people are given a public forum. Kuro5hin took a generally leftist perspective. There are other sites with a similar community-oriented format and size such as FreeRepublic.com that have a quite conservative take on politics. DailyKos is a popular liberal politics-focused site which arose from the voting framework developed for Kuro5hin called Scoop. It is now the eighth largest "blog," loosely defined, according to Technorati[11]. Most of Kuro5hin's politics traffic may have migrated there when the community became unsupportive of discussion.

Stories at Kuro5hin started out as primarily technically oriented but slowly evolved to be more general with a smattering of articles on programming, novel web initiatives and related tech endeavors. The appendix contains several pie charts which demonstrate the primary interests in stories across the seven years of the site's existence. The base of IT-oriented users persists to some extent though many have left since 2003.

As in most online communities the ratio of users to content creators is low. The sites creator has made observations to this effect in the past and the data analyzed confirms this. I cover this aspect further in "Chapter 6: Discussion Concerning the Failure of Community."

Democracy, the Special Sauce

The core interesting aspect of Kuro5hin is its voting queue for stories. Any user who takes a moment to register on the site can both post new stories to the queue and vote on stories already in the queue. Once a story is submitted to the queue it is locked

[11] Technorati Top 100 Blogs List..http://www.technorati.com/pop/blogs/

to give visitors the opportunity to comment on it and grade it equitably. A user has four options for each story in the queue: Post (+1), Post Front Page (+1), Delete (-1), or abstain. Users cannot see the vote until after they have voted and a vote cannot be retracted or changed.

Once the story's score reaches a certain **post threshold** it is posted to the site (to the front page if most of the positive voters specified that option). If it drops below -15 at any point the story is deleted. The post threshold is a malleable value which the site administrator has modified as the site has changed in popularity. It peaked at 145 in May 2001 but was soon readjusted to 95 and was kept at that level from 2002 until mid-2004. It was slowly lowered to 45 by late 2006 when few stories were being posted onto the site from the queue. Eventually, a back door was implemented so that contentious stories could be posted without fully achieving the threshold: if a story has a large amount of discussion and hasn't been deleted for a long period it will automatically post. If it doesn't have a significant amount of discussion it will delete.

As you can imagine, these variables play an important contributing role in the data mining and statistics that will be discussed in later sections. The post threshold is an invisible force on the stories that have been analyzed. If the threshold was out of sync with the population of the site at any given time due to administrative negligence or uncertainty then many articles which would normally have been posted are not. Any article that was deleted is not in the dataset for this analysis. It is unfortunate that some more dynamic threshold was not implemented for it might make analysis easier. If the

deleted articles could be reintroduced to the population we might offset this sampling flaw.

Kuro5hin is not the only site to try open and free submissions. Wikipedia is an obvious example, discussed initially. Everything^2, Digg, Metafilter, Plastic, and the sites discussed above are others. It is interesting to look at these sites as they rose and fell on a common timeline. "Figure 1: Alexa.com Traffic Rank comparisons" displays these fluctuations (one chart for the large scale sites and another for the smaller ones). Everything^2, a project associated with the Slashdot creators was somewhat similar to Wikipedia with a less neutral flavor and was not nearly as prolific. Metafilter is a community site which accepts stories in the format of links and short descriptions from any user, there is no moderation and no threading.

Digg is an extremely popular site that arose in 2003. Its stories are simply links and very short descriptions. It is moderated by voting. Reddit and del.icio.us/popular are two other sites which have arisen on similar ideas in recent years. The central factor behind all three of these communities is a momentum voting algorithm which functions without any manual threshold. It could be said that these sites took a large population from Kuro5hin and that they were the next stage in online collaborative publishing.

Experiments on K5

Anonymous posters were initially allowed but were disabled in September 2000 only nine months into the site's existence.

In mid-2002 the site administrator implemented special features like spell-checking for participants who purchased subscriptions for $3 per month.

In August 2002, Rusty Foster started work to create a non-profit "dedicated to helping support and develop online community and collaborative media." It was called the Collaborative Media Foundation (CMF)[12]. The concept was refined in late November with draft bylaws[13].

The Sky has Fallen

There had always been an attitude that the community was falling apart, even as the site was reaching its peak. As an infrequent participant I dismissed these claims as baseless complaining. These discussions had appeared in other communities online and offline. Now that the data have been analyzed and the community has in some sense died it is clear that the naysayers were correct during at least some large part of the site's arch. There are a few general explanations which may have contributed to the site's downfall.

At some points in 2002 and 2003 the server was overloaded and slow. This state affects the availability of the community, one of the oft-cited requirements of community (Schuler, 1994).

> When faced with malicious users, a large or increasing population, and a high volume of discursive output, online communities react in various ways. Some communities simply disintegrate, unable to cope with external and internal pressures. Others might fracture, with certain members moving their interaction to another social space. (Lackaff, 2005)

Crapflooders began to terrorize the stories. HuSi or "Hulvers site," a diary-only Scoop-based site was an emigrant magnet for diary lovers, DailyKos for politicos, others used

[12] Introducing the Collaborative Media Foundation, http://www.kuro5hin.org/story/2002/8/19/185958/536
[13] Draft of the Collaborative Media Foundation Bylaws, http://www.kuro5hin.org/story/2002/11/27/125038/27

personal blogs. *Lambda The Ultimate* served as a new home for some programming language-oriented users.

Core User Emigration

In addition to the substantial quantitative work discussed below, I used the data behind "Figure 18: Third Quartile post for top users" and "Figure 19: Final post for top users" to generate a list of potential interviewees. I targeted a small number of prolific, high scoring users who listed email addresses in their profile (only one out of ten potential users had these listed) and seemed to have departed the site in 2003. I got responses back from 5 users. I continued a short dialog with them following up on their emails. For several of the other users I was able to find a final post or diary entry which explained their reasoning for departing.

First, users listing addresses were surveyed. Two users left the site because of life pressures unrelated to the site. Another said that his social circle on the site left because blogs were becoming more flexible. A third user considered the downfall in terms of blame: "I think kuro5hin's problems were both administrative and cultural--I think Rusty could have done a lot more to steer things in a different direction, and after a certain point he essentially did very little or nothing on that front. But also a lot of vapid pseudo-intellectualism (wankery) was promoted by the community over what could perhaps have been more substantive discussions." Other users surveyed pointed to the rise of Digg and related sites and considered the Scoop system of voting outmoded.

The diaries surveyed were fairly uniform. Of the twenty highly active users who left in 2003 that I attempted to find, four were purged with their user information removed on request by administrators; one account was locked from an engagement with administrators; three diary entries referenced the decrease in quality and announced exits from the site. The purged information would seem to be a blank vote that quality has decreased so much that individuals no longer wish to have their pseudonyms attached to their comments. One of the diary entries expressed disappointment at the lack of administrative intervention in the community after they had contributed a large amount of money to the funding drive. Another of the diaries was fed up with a particularly tasteless story which had been posted on the site by crap-flooders.

One of the questions I followed up with was intended to explore whether Social Network Analysis would have been useful for the data involved. Online social computing expert Alex Halavais suggested collaborator networks might be mapped as they decreased. One of the users considered that while there were sub-groups within the larger entity of Kuro5hin there were not necessarily cliques.

Rise from the ashes?

There is little in the data to suggest that Kuro5hin will ever return to its previous heights. The emigration has taken its toll on its base of very active and high qualities users.

In a recent and heavily self-referential story entitled “Community Doesn't Scale: Why k5 Is Perpetually Dying,” Rusty, the main administrator, likened[14] the contemporary position of Kuro5hin to the National Public Radio story-telling feature This American Life:

> Groups have peeled off to Digg and Reddit for tech news and random links, HuSi for talking about their cats, dKos and friends for politics. We used to do more of all that stuff. But there's still something here that you don't get anywhere else, and the stats, which have held steady at 2.5-3 million pages a month for a couple years now bear that out.
> My sense is that our niche is original personal stories from so-called "regular people." There are a few other places that do that sort of thing, but none at anything approaching even our seemingly low volume. I would point to the apparently currently-defunct Fray and This American Life as our nearest media relatives, and neither of those does anything more than weekly, if even that.
> Basically, it's a low-volume niche, but we do pretty thoroughly monopolize it.

This analysis is compatible with my own. You can observe the peculiar page view statistic in “Figure 16: Broad dataset figure on logarithmic scale” In counter to Rusty’s weak optimism though, the voting threshold is very sensitive to the population of users, not the population of viewers. Kuro5hin will not be able to serve its new niche as well as it once might had. When it only requires 40 users to post an article as it does in 2007 more low quality stories will get posted.

.

[14] http://www.kuro5hin.org/comments/2007/2/25/184721/191/127#127

Community Site Surroundings (Rank vs. All Websites)

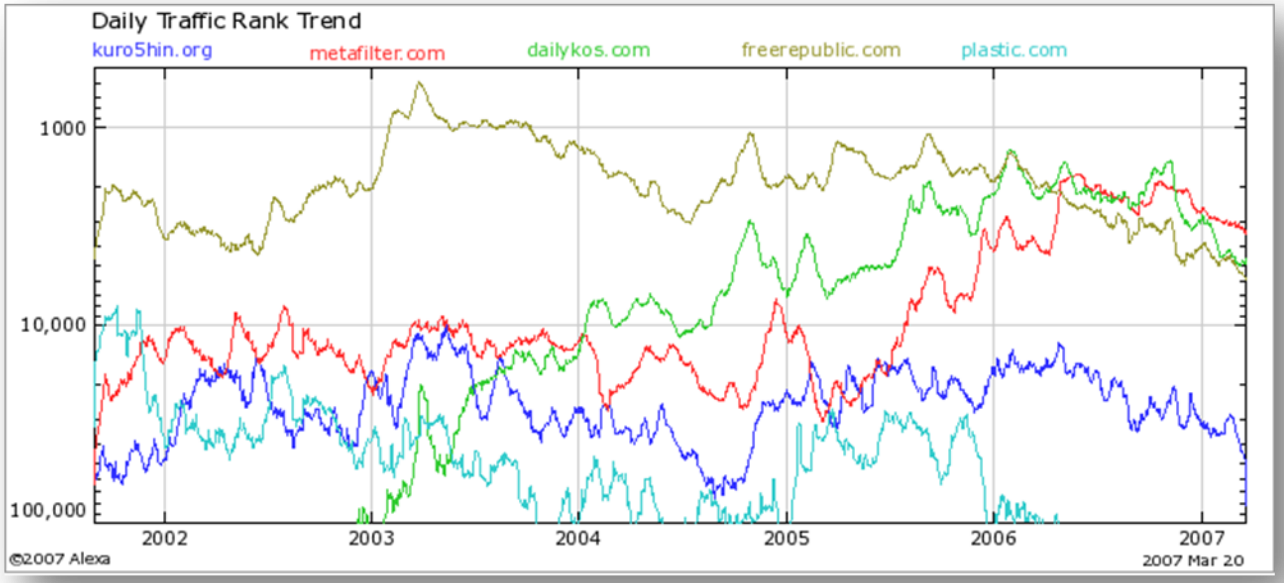

Figure 1: Alexa.com Traffic Rank comparisons

The colors represent different sites between the two graphs (excepting Kuro5hin). Free Republic is only since 2005 receiving competition from Metafilter and DailyKos. Plastic and Kuro5hin have been slowly fading away in recent years. Wikipedia is exploding to the top 10 sites on the Internet while everything^2 is constant at the 10,000th site level. In 2006 Slashdot was outranked by Digg.

Chapter 2: Current Thinking on Internet Community Lifecycle

Virtually Gathered v. Gathered

The Internet now hosts a billion[15] people. "Computer Mediated Communication is communication that takes place between human beings via instrumentality of computers"(Thurlow, Tomic, & Lengel, 2004). A phenomenon resembling community clearly arises from this new nexus of communications. Are these communities bound by CMC comparable to traditional non-virtual communities?

Our Enlightened *Intarweb*

There is clearly a progression of understanding concerning online topics. In the first ten years of the World Wide Web and widespread Internet until about the turn of the century, hype surrounded all the fairly novel concepts. The early prevalence of non-commercial entities (NCEs), Metcalfe's Law and the new media relationship are three examples. The low cost of entry allowed these NCEs to have an online presence comparable to full-fledged corporations. Metcalfe's law, born with Ethernet, suggested that the "value" of connecting networks of computers would be polynomial rather than just additive so N_1 and N_2 together is $N_1 \cdot N_2$ (Latham & Sassen, 2005); and its introduction of "Many-to-Many" interactions was also novel. This Many-to-Many concept is a logical extension of phone technology (One-to-One) and television (One-to-Many). Communications theorist McLuhan's postulate of 'the medium is the message'

[15] Morgan Stanley via http://www.useit.com/alertbox/Internet_growth.html

encouraged this hyperbole: “If, as it is believed, print had made its culture individualistic and linearly thinking and television had brought about a fragmented sensibility and the impulse to ‘amuse ourselves to death’ what would be the profound transformative ‘message’ of online community?” (Feenberg & Bakardjieva, 2004)

The Internet’s noncommercial nature encouraged revolutionary utopic and dystopic ideals. Postmodern utopia was characterized by some as implying a “high-tech social form that can return us to so-called traditional values and intimate personal relationships” (Renninger & Shumar, 2002). Dystopians focused on the digital divide where only the elites have access; the non-committal aspects of virtual community; and the social isolation of the medium (Katz & Rice, 2002). The revolution was not to be, the dot com bubble burst, and reality intervened. The free-flowing venture capitalists’ money petered out. Katz and Rice characterize this original tendency and attempt to mediate it with the concept of “syntopia” in their book on Internet use demographics. In “Virtual Community: No ‘Killer implications’,” Bakardjieva and Feenberg suggest that there has not really been a significant shift in thinking as a result of the Internet.

One of the first researchers to tackle online interaction, Howard Rheingold remarked in 2000:

> As we who observed social cyberspaces gained many years of experience, we came to understand more about the relationship between technological potential and human foibles… I would like to think that Community Informatics marks the beginning of a new era, neither naively utopian nor paralytically critical, based on actual findings by people who have tried to use online media in service of community, then reported on their results. (Gurstein, 2000)

There are many common examples of these early writings which were overoptimistic (Negroponte, 1995; Rheingold, 1993; Turkle, 1995). I will try to view these as most of today's ~~netizens~~ savvy Internet users do: as historical artifacts and reminders of our own limitations.

In addition to this enlightenment, the dominant forms of online community have changed drastically in the past thirty years. The progression includes bulletin board systems (BBSs), Usenet, email listservs, multi-user domains (MUDs), static web communities, dynamic web communities, blogs, and most recently social networking sites. As new forms of the computer mediated communication get more complex the old mediums typically do not die, "they add new practices, often hybridizing, but seldom eclipsing established ones" (Feenberg & Barney, 2004) .

This multiplicity of forms makes it easier to consider researchers' works in two categories: modern and frontier. BBS, MUDs, and static communities would be considered frontier while the other, more contemporary examples would be modern. The discourse can generally be characterized by these patterns as well. I will try to avoid the frontier cases as much as possible. They are less relevant to my primary community, which, though on the cusp of change, is clearly post-bubble.

There are several examples of Kuro5hin-like entities being attempted over the years. Suck.com was a startup and static "community" news site spun off from Wired which has a similar trajectory to K5 (Sharkey, 2005). More recently in June 2005, a startup called Bayosphere attempted to make a news site by, for and of the San Francisco Bay Area which sounds eerily like Kuro5hin with a physical focus (perhaps

imagined as successful seed city because of earlier projects that grew there—the Well and Craigslist) (Grubisich, 2006). Unfortunately it was not to be—only 400 articles were published on the site in the seven months it was online, about two articles a day, lower than Kuro5hin is even now.

Real™ Community

The fundamental issue at hand for my project is "What does the online community do for its constituents and why might they spend their time at one?" These questions self-select against the early intellectual debate concerning the value of online community and refocus the inquiry on the prototypical user. It is a functional analysis.

> The term *social relationship* will be used to denote the behavior of a plurality of actors in so far as, in its meaningful content, the action of each takes account of that of the others and is oriented in these terms. The social relationship thus consists entirely and exclusively in the existence of a probability that there will be, in some meaningful understandable sense, a course of social action. For purposes of definition there is no attempt to specify the basis of this probability. (Weber & Parsons, 1947)

Essentially real culture is culture that is not packaged for mass consumption; rather it is something special that arises out of human interaction. Arendt calls this durable culture. "The crisis in culture and education is that now education is transmitting things that are not durable" (Arendt, 1958). Human beings come and go but the world endures. Combining this idea of a social relationship with this durability gives us something resembling Social Capital which is concrete—durable and nondurable social relationships. This concrete definition is important as it hints at the true source of the quality which arises from the consultation of the user population.

Democracy is the interesting element of the communities that I studied. It is commonly associated with durable social relationships. These Internet communities are communities of interest so when the topic fades, perhaps so does the community. A durable community is perhaps the most real community, regardless of whether it exists online or off. This said it seems probable that, at least from instant to instant an online community should be able to sustain the durable classification and provide its users with the plentiful benefits of democracy.

Basic Concepts Underlying "Community"

Community is a tighter and more cohesive social entity relative to a society that is bound together by a “unity of will” – Ferdinand Tönnies, German sociologist, 1887. (Tönnies)	Basic sociological qualifications: 1. Membership 2. influence 3. integration and fulfillment of needs 4. shared emotional connection (‘shared identity’) (McMillan and Chavis 1986)
"Web communities happen when users are given *tools* to use their *voice* in a *public* and *immediate* way, forming *intimate relationships* over *time*." (Powazek & Safari Tech Books Online, 2002)	Social Capital: "the collective value of all social networks (who people know) and the inclinations that arise from these networks to do things for each other (norms of reciprocity)." (Putnam, 2000)
“Social identity has an important psychological function. It can contribute to our self-esteem by allowing us to assume we possess the positively evaluated characteristics that are stereotypically attributed to the groups with which we are associated”(Sherman, 2003)	

A few distinctions of online communities
• Individual can completely tailor personal communities; (Renninger & Shumar, 2002) • Archives can allow nonlinear, threaded interaction; • Strength of Weak Ties is amplified—people who are on the fringes of a social network usually have the most to offer in terms of new information; (Granovetter, 1983; Whittaker, Terveen, Hill, & Cherny, 1998) • “The ability to come to identify with a group online, and the support to do so actually provides a scaffold for a different and enhanced sense of possibility for individuals” (Renninger & Shumar, 2002).

Table 1

The Myth of Community

"Reality is that which, when you stop believing in it, doesn't go away." (Dick, 1968)

Benedict Anderson, who studied the concept of the "nation," claimed in 1983 that "All communities larger than primordial villages of face-to-face contact (and perhaps even these) are imagined." Online communities are unabashedly imagined: the rest of Anderson's claim states that the "style in which any given community is imagined distinguishes it from any other community" (Feenberg & Bakardjieva, 2004). "The fact that virtual communities are defined by contents for which community has an interest is one of the reasons that critics tend to see virtual communities as something other than community. Participants' connections to community are both cognitive and affective, rather than simply spatial and temporal" (Renninger & Shumar, 2002). Lurkers, or passive participants are outside the affect and are parasitic on the cognition. Where all communities are social constructions, Internet communities consist purely of artificial connections forged and mediated by computers. This doesn't mean that they are not meaningful.

In most recent discourse, Putnam, the inventor of social capital, is not ready[16] to commit to most online sites' communityhood (withholding full marks for his social capital and shared identity). Putnam examined the online classifieds "community" Craigslist discussed in Chapter 6. He and his co-authors think it is still an open question.

[16] "We are in no position to judge authoritatively how much real community building and social-capital building happen on the Web. Although we looked hard for candidates for a case study of online social capital, our investigations usually turned up less solid evidence of new social capital than we had hoped. Some apparently promising sites were in decline; some turned out to be commercial ventures with a veneer of community vocabulary; most offered no clear evidence of members' building the relationships of trust and reciprocity that we understand to be central to social capital." (Putnam, Feldstein, & Cohen, 2003)

They are unsure mostly because of frequent declines in the communities of the Internet (2003). Meijas agrees with Putnam in a limited way: he argues that we are too indefinite with our value judgments of our own online interactions. We either say they are only virtual (faux) or exaggerate their importance past reality (Mejias, 2005). Social network researcher Hawthornthwaite suggests that the lack of “social presence“ and the feeling of “being there” damns the ‘community perspective’(Renninger & Shumar, 2002; Short, Williams, & Christie, 1976). It is too easy for onetime community members to stop believing in the online community and make it disappear. Social psychologist Sherman (Professor Emeritus at Miami) disagrees:

> First, despite the apparently impoverished text-based nature of most forms of CMC, people do form impressions of each other and they do develop strong interpersonal relationships online. Second, CMC may foster perceptions that are more extreme than in face-to-face situations, but the positive or negative direction of the effect may depend on factors external to the medium itself. (Sherman, 2003)

Other sociological voices like Bakardjieva dismiss this general question of real versus virtual; they find it easier to approach the topic from the dichotomy of human interaction and commerce (Feenberg & Barney, 2004) .

Benkler takes a similar view but focuses on the Internet communities of the “commons” (Benkler, 2006; Raymond, 1999). A "commons" is a place where everyone has equal and complete access to all the resources. People are naturally creative. They want to make. For centuries, the industrial economy required significant startup capital to create anything. The Internet, however, affords “capital capacity necessary to do so; if not alone, then at least in cooperation with other individuals acting for complementary reasons” (p. 6). Due to the lower barrier to entry, people can easily

draw resources from online communities and often enough find it convenient to add useful things back in.

"The idea of virtual community is indeed a powerful myth playing on the genuine desire we have to control our lives and be a part of a larger social whole that provides emotional and intellectual support"(Feenberg & Barney, 2004). As a student in the Western College Program community, I know that there are many tangible, positive aspects of a real live community but a significant part of my experience with Internet communities suggests that they can be powerful as well. All sorts of communities are subject to moments when they seem ephemeral and other moments when they seem sure as silver. Myths are stories which are too important to be ignored and if nothing else, all these researchers agree that these semi-coherent collections of social beings must be addressed.

The observer and the observed: Lurking, the participants and authority

Social Networking theory says that you have a number of people with whom you associate at different levels. Those people also have their own set of associations. Within the investigations of computer mediated communications (CMCs), the extension of Social Networking theory to cover people's interactions online was quickly obvious. "We can define community based on what we do with others, rather than where we live with others in terms of the social networks we maintain." (Renninger & Shumar, 2002). The few people closest to you, your friends and immediate family, are nearest in your network. The further you go from this core, the greater the number and the weaker the

ties in the network. In a landmark article, Granovetter pointed out that the people nearer to the **periphery** are the most valuable in terms of knowledge—the people nearest you tend to think just like you (Granovetter, 1983).

Internet communities offer a unique opportunity that exploits this counterintuitive conclusion. They allow the users to break out of their everyday core network and into a nearly infinite new set of ties.

"The Kindness of Strangers" connects this theory directly to organizations and computer mediated communication (Constant, Sproull, & Kiesler, 1996). The article talks about how informed employees of a corporation share computer knowledge with less informed employees even when there are few clear benefits to the informed employees. There is a lack of easily identifiable reciprocity in technical support across the companies examined in the study. Employees were exploiting weak connections.

In a community of interest like Kuro5hin, infusing new blood into a community is important for just the reasons that talking to people you know least is important. There is value found in community besides that of forming close friendships. All the users are gaining valuable information from people they can now associate with and would not have been able to before. The lurkers and the people more distant from you have more information than you that is different from your own. This theory defines the community of discourse. We have to remember, though, in communities like Kuro5hin, there are always people who are sitting "outside" watching the authors write the articles.

With any web community, there is a large elephant in the room: the multitude of invisible, noncontributing observers. The Lurkers. I will quantify them in later sections but a community online is often considered as the engaged few who perform for the unengaged many. Even these lurkers often identify with the community and obviously gain information from it, even if they are passive. It is easy to forget about these others but they do play an important role in the community (Nielsen, 2006; Takahashi, Fujimoto, & Yamasaki, 2003). In one conference presentation Lee, Chen and Jiang "uncovered [the lurkers'] process of negotiability and identification from a virtual ethnographic approach." (Lee, Chen, & Jiang, 2006)

As Benkler explains, there are tangible outputs from many online communities— computer code, good articles, or FAQs. The participants contribute to create these outputs. At one level, the lurking hordes are simple freeloading. At another level they are actively processing the information, taking it into their other communities and social networks in which they may be more active. The participants do not necessarily think about how much they are affecting the lurkers. I draw these points back to Renninger's. In the heated debate above she advocated that online communities be appreciated primarily for their information's value (the "Cognitive" connection). My goal with my creative research that follows is to tap and measure this information flow to the best ability of the tools available.

Chapter 3: Rationale

Hypothesis

In this chapter I briefly present my expectations given my methods. Again, I suggest that it is possible to evaluate an Internet community for Intellectual quality. This is a favored notion in academia but it must be bridged with the simple curiosity of intelligent people in order to have some useful way of examining such public unmoderated discussions. Quality can be defined operationally: A post is of high quality—intellectually thorough in the context of Internet community if topical and seems to include high grade word structure thinking and assuming that someone is not trying to subvert the measure. Signal overlaps the domain of rigor in a qualitative sense.

These are my assertions for the creative assessment portion of this project:

- All seven readability tests will be statistically similar.
- The readability scores will be highest in 2001-2002 corresponding to the story activity on the site.
- Latent Semantic Analysis will predict ratings with similarity given a reasonable topic space.
- Many important users key to keeping the sites intellectual rigor sustained left the site after 2002.
- The readability scores correspond in some weak way to the rigor of discussion.

The community ethos contains value. Maintaining this ethos requires constant administrative monitoring.

Value is important in searching. It is trivially easy to write a good search algorithm based on a "hash" table as an index. Prioritizing the results in a way that is useful to users is **very** difficult. Google, for instance has made its fortune on deciding what is valuable for people to see. It uses an algorithm called PageRank and a variety of undocumented tweaks to determine which search result containing your terms should show up at the top of the results. In this context, value of documents has been deeply explored. The combination of measures I propose might be able to extend this notion to contributions online.

PageRank primarily infers quality by noting the sites linking to the site at hand. There are few cases of systemic linking inside the community space that this project examines so this approach may not be appropriate. Novel techniques such as the ones I attempt have potential.

Chapter 4: Methods

Signal v. Noise

The goal of data mining is generally to filter the useful stuff from the useless stuff. In my case, the useful stuff are comments and stories which are useful to people. These data brings the question of quality of interactions as that is what is important to users. It is clearly a tough nut to crack for a computer since it is inherently qualitative. I have coarsely identified four categories of estimating quality of my primary community:

Human	**Readability**	**Tangentiality**	**Semantic Tests**
GOAL: Comprehension	Flesch test (Word's classic readability test)	Custom Wordnet test (Lambrecht, 1994; Marcu, 2000)	Latent Semantic Analysis(P. W. Foltz, Kintsch, & Landauer, 1998; T. K. Landauer & Dumais, 1997; T. K. Landauer, Foltz, & Laham, 1998)
Heterarchical Ratings (inherent)			
Votes (inherent)	Dale-Chall (Chall & Dale, 1995)		
User Interviews	Cognitive Readability(Chall & Dale, 1995)		

Table 2

Pyle defines 10 commandments of data mining

> Select clearly defined problems that will yield tangible benefits; Specify the required solution; Define how the solution delivered is going to be used.; Understand as much as possible about the problem and the data set (the domain) ; Let the problem drive the modeling (i.e. tool selection, data preparation); Stipulate assumptions; Refine the model iteratively; Make the model as simple as possible- but no simpler; Define instability in the model (critical areas and ranges in the data set where the models produces low confidence predictions/insights) (Ye, 2003)

It is hard to refute many of these since they are valid guidelines to keep in mind as I discuss the other material in this section and they set the tone of the section. That article continues with more specifically detailed guidelines as well.

A reasonable metaphor for what I am trying to do is the way that market trends are used in politics. The patterns are a tool of depersonalization which is analyzed by experts and politicians and eventually converted back into something that affects people often in the form of laws. For the moment I will defer to these presumably impersonal analytical schemes. Once the results are in I can express my personal analytical inclinations based on the academic context from my other subtopics.

I have considered that criteria for "quality" of human discourse are difficult (impossible?) to derive passively. Perhaps if it was possible to address the entire population and ask them some questions I could do better. Then, I would be able to add some test questions that would trap for misleading, useless prose but this is outside the scope of my project. I would like to note that there is some human input on quality available for analysis—the ratings on comments and votes on stories can be used in a number of ways. Comments or stories that are rated highly might be used to bias the other metrics. If the average (or STD, etc.) of the human metrics numbers changes over time I might even have a decent basis for the quality over time graph itself.

Readability

There are two sorts of readability metrics today—classical semantic and cognitive. Classical semantic readability has been around since the first word lists in the late 1920s (Chall & Dale, 1995). Cognitive methods for assessing readability involve our

understanding of the difficulties of language. Lacking a computer framework for the cognitive approach, it is generally out of this project's scope.

Chall points to research that indicates readability metrics are not useful for writing or rewriting in general (Chall & Dale, 1995). This analysis might imply that they are not an ideal judge of quality. There is also the question of the cognitive and structural features of the text like how many long-term memory retrievals are necessary to understand something. These techniques began to be studied by Kintsch in the late seventies ().

In one study, Kintsch analyzed the campaign speeches of Eisenhower and Stevens (then opponents in a presidential race). Stevens' speech was generally considered to be more difficult. He found that Stevens' oratory required the reader to make three "networks" of connections to understand the meaning whereas Eisenhower only required one. This contrasts with the Dale-Chall classic test which suggested that the talks were of about equal difficulty.

If we were trying to guess at the quality of the article itself we may have data but selected, cherry-picked data without context which might not be as representative as we might like:

> Our predictions about the effects of demographics on conversation strategy were largely confirmed but we found disconfirming evidence about the relations between conversational strategy and interactivity. Contrary to our expectations, both cross-posting and short messages promote interactivity. (Whittaker et al., 1998)

This finding might have negative implications for any classical readability metrics where longer length sometimes means higher score. The classic readability tests also fail badly when trying to assess mathematical notation.

Tangentiality

"WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets."[17] To restate the official description, the primary way that words are defined in WordNet is by their synonyms. This is a useful way to organize things for a computer because dictionaries can be represented as a graph; all graph search and utility algorithms can be used for processing. The Ruby language has libraries to use WordNet natively. Topic words are demarcated in language. Languages like Japanese make it even more explicit but words like 'regarding', 'given', and 'for' are common indicators of topic in English. When I designed this metric I might have used some of the text processing which have chapters on sentence "topic" (Marcu, 2000) and general sentence structure (Lambrecht, 1994) to see if there are further patterns which may be easily accessible and exploitable.

As I researched I did find an article which suggested an algorithm for query to document comparison using WordNet (Varelas, Voutsakis, Raftopoulou, Petrakis, & Milios, 2005) but little on document to document[18] comparison. Unfortunately in the

[17] http://wordnet.princeton.edu/
[18] That is, Story to Comment where both are documents.

discussion section of this article the authors state that even such query-to-document uses of WordNet have historically been unsuccessful. I also found studies of Latent Semantic Analysis that suggested that its broader methods were significantly more successful than simple synonymy in correctly comparing meaning. I ultimately decided to focus on other methods. LSA could do essentially the same semantic or meaning-based story-to-comment comparison I was looking for with less time consuming original research and better accuracy.

Semantic Tests

Semantic Tests are sophisticated packages which might actually provide a better tangentiality measure than the one that I propose above. Due to time constraints on the project these were not completely explored experimentally. Definition and theoretical consideration of this class of techniques is provided in the "Semantic Tests" section of the Discussion chapter.

Ethical Considerations

As an Internet researcher I had to make certain decisions based on feasibility and ethics. My subjects included people who posted to the Internet with the knowledge that it might be downloaded by others on the Internet. My use of their data is only passive so it bears little on their personal lives. I did not include the semiprivate sections of the site like the diaries of individuals and avoided any methods which might turn up any incidental personally damaging information in my dataset. This is in line with other similar studies—"the nature of my research, the nature of the data collected and my method of data collection post minimal risk to the the case study community or their

members" (Lackaff, 2005). I ultimately used methods similar to Google's searching and cacheing engine or Alexa's Internet Archive. I have no intention to profit from this endeavor. A best effort was made to spread the server load over several weeks to prevent the large volume from having an effect on regular service.

Procedure

Speculative Groundwork

A side project was undertaken initially. I realized that I would be able to graph out a rough users-over-time graph with a little speculation and almost no data in my database. Since the site used counting, relatively continuous numeric user identifiers it was easy to jump directly to user #40,000 or #80, that is the forty-thousandth registered user and the eightieth registered user respectably. I could then see when their first comment was made. Initially I was skeptical that this first-post date would be reflective of their account creation date but after verifying that the surrounding 5-8 accounts had similar dates I followed through with the technique. I collected 32 samples from throughout the set of user IDs, checking the surrounding IDs to corroborate their true creation with some rough degree of confidence for each of these samples. I then plotted these points to get a general idea of the user base growth timeline of the site. I produced "Figure 10: Registered users over time" and "Figure 11: New accounts per day" from these data. This turned out to be an effective quick and dirty method to divine some valuable trends without any screen scraping early on. It has some reusability among sites like Slashdot which have numeric and counting user IDs.

Surveying

The first main objective was to draw down the indexes of all the stories posted to the site since its beginning. “Figure 4: Data summary” illustrates the data involved. The site is based on a back-end database and software that accesses it so this old content was just as accessible as the newer material. Initially I set up an account and configured it so that the content I was looking for was displayed and simplified in a way which is most convenient to parse. I used the HTTP cookie left over from my browser and downloaded some 250 pages containing 24 stories each. Although the Kuro5hin.org site does have a diary section as discussed above, I neglected those in favor of the full fledged voted-and-posted stories. After each download I used htmltidy to clean the HTML into XML format. I then “scraped” the unnecessary portions of these indexes off keeping only the interesting information: comments per story, words per story, titles, URLs, date, and author. In modern sites this scraping process might be less laborious thanks to newly established XHTML/CSS standards but the kuro5hin site was statically designed with HTML tables. Hand-optimized XPATH queries, generated using Visual XPATH[19] were executed in C# for this part of the processing. It took many hours of tweaking to account for all of the data on all of the pages previously downloaded and ultimately even then there were issues that had to be resolved by hand and reimported. This resulted in 60mb of raw data and 4mb in an organized spreadsheet. Minor changes in a few aberrant posts were corrected in this spreadsheet and the contents were fed into a MySQL database.

[19] Visual XPath http://weblogs.asp.net/nleghari/articles/27951.aspx

Materials Acquisition

The next step was to download the actual content of the site. Again, I configured a user so that the comments were in the best layout for scraping. The URLs that were trimmed from the indexes were, one by one, downloaded with a Ruby script, separating each request by a minute or two. The script only downloaded during off-peak hours (2am-noon). Surprisingly, some of the stories with comments were approaching 1mb (> 770kb) in raw size. It was only polite to be cautious with server load and bandwidth, especially in such volume (6600 articles). In total, the article base took a week and a half to download and used 1.3gb of storage. With preliminary processing it appeared that 1.0gb of this is comment content. After each download I used *htmltidy* to clean the HTML into XML format.

Construction

The following step was loading these new data into the existing MySQL database. This was also accomplished with Ruby script. The Hpricot library provided fast processing of XPATH in Ruby—crucial with the comment volume involved. MySQL 5.0 stored functions were also lightly used to normalize comment ratings across the seven years of the site since the scale switched from 1-5 range to 1-3 range in Oct 2003. Many concessions had to be made for quirks in the data so I had to manually import the stories (and contained comments) in sets of about a thousand. I chose 5 as the normalized maximum rating. Processing could proceed in parallel on my dual-core machine but it ultimately took about five hours. I noted that 596,872 comments were successfully imported in the database from 6,368 stories. 289 stories scattered

throughout the sample failed to import for various reasons including malformed HTML which didn't successfully convert into XML. These were discarded.

Once these data were in the database it was easy to query for individual commenters' post-count, their first and last comment date, their average (human, site defined) comment rating and their average comment date. The tally allowed me to contact the most prolific posters with high ratings who left the site early on to get more qualitative and retrospective justifications for their departure. Many on-site diaries of these individuals also proved useful. There loosely seemed to have been about 500 users which were especially valuable to the community with high comment post counts and high ratings. I created a column which combines these values (post count x (rating/5)) with story post count to produce a rough score.

One aspect that I considered crucial to my hypothesis was the dates that people were participating in the community. To this end I installed the C based MySQL User Defined Function library to calculate the median[20]. I then modified it to also calculate the first and third quartile. Through a simple query with its output exported to Excel, this allowed the production of box plots for community members' dates of participation. The fifty highest scoring users' box plots are shown in Figure 17. The data were quite overwhelming and will be discussed further in the results.

The rest of the data scraped from the stories included the introduction to the story, the actual contents from the story and all the comments in that story (numbering typically between 100 and 200 but in a few isolated cases above 1000, see "Figure 13:

[20] http://mysql-udf.sourceforge.net/

Word count in stories"). Storing the hierarchical, threaded, parent/child data in a relational database was a bit tricky but ultimately I settled on an Adjacency List Model instead of a Nested Set Model[21]. I didn't anticipate using those data extensively enough to make the initial load time worth the implementation difficulty tradeoff.

Processing

Finally, the processing of the actual text began. The first textual tests involved the standard unix tool style from the package style & diction. The tool produced seven Classical semantic readability metrics (Kincaid, Automated Readability Index, Flesch Ease, Fog Index, Smog, Lix and Coleman) among a host of other measures including passive-voice use and verb use. I ran this tool on every story greater than 300 words and then on every comment greater than 30 words and updated the database so that every record was tagged with each of the seven results. I used the Dale-Chall algorithm I discussed above as a prototype Classical semantic readability test. It is similar to these but was not utilized due to time constraints.

Eyeing Lurkers

A side project was proposed at this point to gather visitor and hit information from a set of public *webalizer* web server log pages. These pages specify daily load for every day starting when the site had only 33 users (February 20, 2000). It noted exactly how many unique visits occurred on the site but only through March 2001. It also noted exactly how many hits occurred to the web server on any given day of the site's existence. Additionally it listed a number of irrelevant statistics such as data sent from

[21] http://dev.mysql.com/tech-resources/articles/hierarchical-data.html

the server. These were discarded. I expected the hits and visitor data to be useful when considering volume over time. I used my tool from the first scraping step to quickly generate a spreadsheet with the dates and values. Eventually I imported this data into the SQL database. The results of this processing are shown in “Figure 16: Broad dataset figure on logarithmic scale” listed under hits and visits.

Chapter 5: Results

Stories Volume

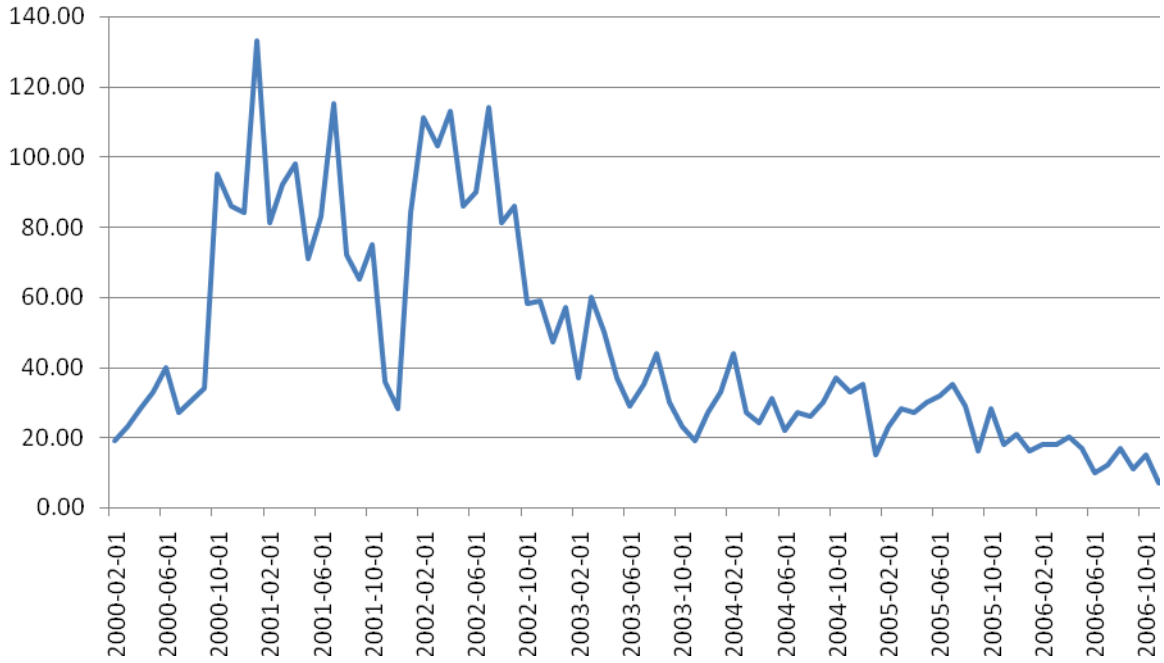

Figure 2. Story volume over time

The first set of data that was returned was the usage plots. Simply looking at the usage graphs, as you might expect, it is impossible to distinguish between interest and quality. More visitors may mean higher quality but it is hard to be certain. The strange dip in August 2000 and December 2001 reflects long bouts of downtime.

Average Coleman Readability Test Result of Even Month Stories vs. Time

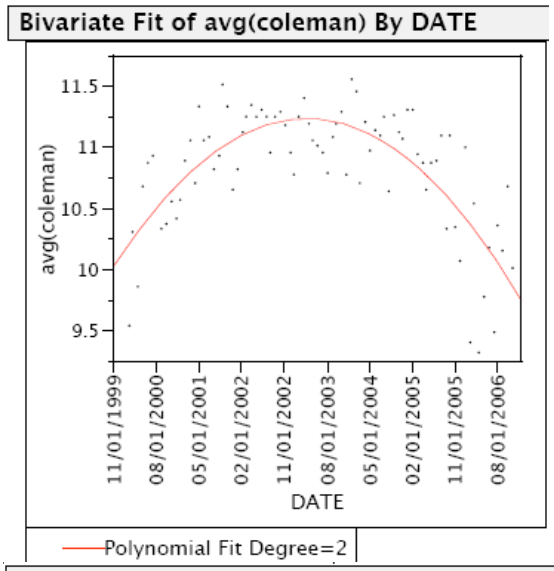

Figure 3. Average Coleman Readability vs. time
Each of the readability tests showed a relatively clear trend over time when using the average. The Coleman test showed the strongest potential for predicting the signal vs. noise.

Polynomial Fit Degree=2

avg(coleman) = 16.808525 - 1.78e-9 DATE - 1.1e-16 (DATE- 3.1e+9)^2

Summary of Fit

RSquare	0.566234
RSquare Adj	0.555112
Root Mean Square Error	0.33619
Mean of Response	10.8296
Observations (or Sum Wgts)	81

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	11.508111	5.75406	50.9102
Error	78	8.815841	0.11302	Prob > F
C. Total	80	20.323951		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	16.808525	1.899595	8.85	<.0001
DATE	-1.78e-9	6e-10	-2.94	0.0043
(DATE- 3.1e+9)^2	-1.1e-16	1e-17	-9.70	<.0001

As of December 5, 2006 there were 69,603 registered member of which an unknown number are active (probably in the several hundred). More figures that were produced in the creative portion of this project can be found in Appendix II. These include graphs of the size and activity of the user-base. The exploratory tests for Latent Semantic Analysis did not produce any figures.

Chapter 6: Discussion Concerning the Failure of Community

"We're trying to figure out how to run the site as a commons, yet avoid the tragedy of the commons," Craig of Craigslist explains[22]. (Putnam et al., 2003)

In his seminal article on "The Tragedy of the Commons," Garret Hardin began with a note about nuclear engagement and how physicists concluded there would never be a technical solution to the problem of the arms race. (Hardin, 1968) One of his primary ideas in the article is that for many problems there are no technical solutions. The other is more common knowledge: people tend to maximize their use of an unmoderated resource. The significant solution he suggested was "Mutual Coercion Mutually Agreed Upon" where the relevant parties get together and make laws. There must be rules set down to guide participants and these cannot be completely technical. An intelligent experimenter can find a way around almost any software wall in a community but if many people agree on a rule then heterarchal, public moderation can hold order.

So far this project has established that Kuro5hin is unlikely to return to prominence. To any honest observer its standards have dropped in rigor and topic. Craigslist, the seventh largest commerce site[23] on the Internet, a free metropolitan classified advertisements service, has thus far stood the test of time where K5 has not. The niches are very different, but the comparison is still appropriate.

---

[22] "We still have a ways to go. There's always going to be something." And "the culture of trust we've built is a really big deal. We have to re-earn that every day."

[23] http://www.marketwatch.com/news/story/story.aspx?guid=%7B11A7B265-1898-49BB-9D4F-7D414D91F8DC%7D "Craigslist thumbs nose at Wall Street." MarketWatch.

Community Supports

Viable Community

In one of the earliest articles on CMC communities Schuler points out: “the issue of attitudes and values—the politics of community networks—makes participation in the community network development important” (1994).

Early conception of minimum technical supports for online community:

1. **Bounding**: forming closed online groups;
2. **Tracking**: Listing how far each participant has read in community discussions;
3. **Archiving**: maintaining accessible record of community discussion;
4. **Warranting**: ensuring stable and (most of the time) genuine participant identities.(Feenberg & Barney, 2004)

It was soon clear that in Usenet, at least even these limited supports had not been implemented fully—people could still pull off anonymity. In Mass Interaction “common ground model must be modified to incorporate notions of weak ties (Constant et al., 1996; Granovetter, 1983) and communication overload (Whittaker et al., 1998). This conclusion emphasizes the lack of the minimum technical supports of tracking and identity. Other supports include the administrative duties such as the deletion of problem users, the maintenance of the site software and an open feedback connection to the community. An administrator of such a community needs to stand by the community. These tasks that are required of them become more onerous as the site grows in population. If it does not scale so that duties can be shared with the growing community then the growth will outstrip the capacity of the administrators’ ability to handle it. Wikipedia deals with this elegantly. There are more administrators as there are more users. Kuro5hin never really involve d more than a handful of editors and when too many of them got bored or frustrated with the duty, collapse was imminent.

Software and its Wetwares

People in online communities today expect and require certain supports. Theories concerning these supports have evolved since the original four. Today, some sort of trust, rollback, or rating system is required for quality control. “Empirical studies show that in practice, interacting users appropriate the technology as members of a particular social groups with particular goals in mind. In that context they discover and enact new affordances not always deducible from obvious technical features.” (Renninger & Shumar, 2002) Sure, much community building is nontechnical but some of it can be technical. Even Wikipedia, an implementation of “the simplest database possible” concept of wiki, has a revision history to purge vandalism. Usually communities have some members who have special privileges: in my primary community there were editors who might delete someone’s comment if they had second thoughts about posting it. These community police are a minority though; mostly the users moderate themselves through voting and ratings. Self-moderation is much more efficient for a community of any significant size.

Scale

Participant Inequality

A prime issue of Internet communities and to an extent any community is Participation Inequality. The leader of my primary community has released statistics that indicate for every 200 readers there are 20 casual commenters and voters and 2 active posters and story writers. Nielsen provides insight into the larger Internet. He first cites research by Will Hill into user participation. That research indicates 90% of any user

participation are “lurkers” or passive readers, 9% consist of users that contribute occasionally and 1% are users that make most contributions. Nielsen then investigates Weblogs and Wikipedia (Nielsen, 2006). There have also been investigations into Usenet where similar observations were made (Whittaker et al., 1998). Nielsen concludes that this is an insurmountable problem and should be addressed simply by not treating it as a problem and by lowering the barrier to entry as much as possible.

As you can imagine, this small participant group might be in the position to run the system. In Digital Democracy, Hague lauds the open, shared aspects of the Internet compared to mass media and talks about how they will aid open democracy. He also speaks of how democracy is always a see-saw between control of people at large and elites. His goal is to suggest certain forms of productive organization online (Hague & Loader, 1999). Democratic elements like the voting of stories in Kuro5hin are a mix of consensus and conflict—two strategies for community development (Chaskin, 2001). On that site, editorial comments which are submitted before the story goes “live” encourage improvement and regular comments can be encouraging or challenging.

Durability: Does Community Scale?

“A community of things which gathers men together and relates them to each other depends entirely on permanence” (Arendt, 1958).

In *The Vanishing Table, or Community in a World That is No World* Darin Barney references Arendt’s statement above (Feenberg & Barney, 2004). He goes on to connect the durability of community to something more, a shared world or shared history. In the first several years Kuro5hin members had a world. They were computer professionals or

those that aspired to be. They were interested in events on the Internet and otherwise. The site's motto, "technology and culture from the trenches," held the users together. Every aspect of the Scoop system worked for the user base—diary pages, comment posting, even advertisements hosted community forums.

What makes mass society so difficult to bear is not the number of people involved, or at least not primarily, but the fact that the world between them has lost its power to gather them together, to relate and to separate them from others. The weirdness of the situation resembles a spiritualistic séance where a number of people gathered around a table might suddenly, through some magic trick, see the table vanish from their midst, so that two persons sitting opposite each other were no longer separated but also would be entirely unrelated to each other by anything tangible. (Arendt, 1958).

As time passed, as information overload from crapflooders took hold, as servers ran glacially, and as administrators lost the spark and did not add modern supporting features across the site like RSS[24], the table somehow vanished. Work on stories became labor. The prolific users were not so much sharing among interested compatriots as apathetic strangers who were intent on overloading the discussion. All of these components detracted from the permanence that the Scoop software provided.

Semantic Tests

Latent Semantic Analysis (LSA), sometimes called Latent Semantic Indexing (LSI) is a theory and method for extracting and representing the contextual-usage meaning of

[24] "Really Simple Syndication"is a way to easily congregate many sources of news easily.

words by statistical computations applied to a large corpus of text. It is related to neural net models but is based on singular value decomposition[25]. In their canonical introduction, the Latent Semantic Analysis (LSA) creators state that it "accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay" {64 Landaur, Thomas} .

As I was planning, I was set on using a Wikipedia-trained Latent Semantic Analysis model to compare stories versus comments. Wikipedia is ideal because it shares the understanding of Internet culture and discussion with Kuro5hin. It is still a feasible strategy but it turned out to be significantly more involved that it seemed. For most data mining models you have training and testing sets of data. The model should hold quite well on the training data and fairly well on the testing data in order to be a valuable model. As LSA works, though, it is relatively easy to use one of the many existing "topic spaces," as they are called but difficult to create your own. Chapter 2 in the *Handbook of Latent Semantic Analysis* describes each step and the array of different tools available to accomplish each step: parsing, single value decomposition, and vector operations. It would have taken several extra weeks to accomplish each step here (T. K. Landauer, 2006).

There was a second topic space problem: essay grading techniques with LSA require the topic space to be especially well versed in the specific domain of the essays being grades (P. W. Foltz et al., 1999).

[25] "Single value composition, a mathematical matrix decomposition technique closely akin to factor analysis that is applicable to text corpora approaching the volume of relevant language experience by people."

In lieu of these ideal topic space setbacks I performed some preliminary comparisons with LSA intended primarily to test the water for future work. I used the Touchstone Applied Science Associates (TASA) topic space on the Official LSA Website[26]. It is trained on some 70,000 documents that college freshmen might have been expected to have read through that point in their lives. One deficiency is the TASA corpus is missing such important words as Internet.

Experiments

In Kuro5hin, the top 800 users produced half of the posts on the site. My arbitrary participation score formula suggested 250 users of these were particularly active. Using my data I am able to make some clearer observations concerning my graph. Initially I expected that the quality would clearly drop after June 2003.

In response to my assertions:

- All seven readability tests will be statistically similar.

All the tests were somewhat similar but the Coleman test, the test with the best fit for the monthly data, had the lowest R^2 (typically around 0.5) when compared in a bivariate test against the other readability tests. The other tests compared among themselves with high R^2 between 0.79 (Kincaid & Smog) and 0.96.

- The readability scores will be highest in 2001-2002 corresponding to the story activity on the site.

It turned out that the readability scores peaked in April-June 2003 in almost all of the readability tests.

[26] LSA Website. http://lsa.colorado.edu

- Latent Semantic Analysis will predict ratings with similarity given a reasonable topic space.

It is hard to make any certain calls from the limited testing performed with LSA. It did subjectively seem that comments on stories with a decent amount of content related to the story were given higher similarity scores. This finding turned out to depend quite a bit on whether the important keywords were in the topic space of the online site which was very limited in its coverage concerning Internet and cultural jargon.

- Many important users key to keeping the site's intellectual rigor sustained left the site after 2002.

In analysis I have discovered that the causality of this statement is very difficult to approach. My best effort is made through the figures in "The Corpses of Pseudonyms" found in the Appendix II and the user surveys I performed in "Core User Emigration" found in Chapter 2.

- The readability scores correspond in some weak way to the rigor of discussion.

It is somewhat strange that the readability scores did not peak until after the story volume had already dropped quite precipitously. The curve is very clear though. Such a high R^2 indicates that there might be a trend to build further analysis upon.

I did make some effort to check the regression by breaking the data into even and odd months and similar trends persisted. I considered testing readability against the hit count but as you might observe in "Figure 16: Broad dataset figure on logarithmic scale," the hit count has been constant since 2002 so it is probable that such a test would yield little. This data were originally intended to find Lurkers but this turned out more difficult than it seemed. The administrator-managed threshold (discussed in Chapter 1) and the population of voting users also together may be changes in statistical populations and might require distinct regression analysis for every major change.

Conclusion

Each analysis for this project has been more intense than I expected and I have finished only a subset of the goals that I originally set out to accomplish. This is acceptable—such problems tend to balloon when actually explored. The biggest avenue which has been thoroughly researched but not pursued is Latent Semantic Analysis. I was disappointed by the lack of tools to easily run the analysis—although there have been hundreds of studies on the topic for research such as this, few actually create a topic space for their data. Though I must submit this project in time for a deadline I plan to continue pursuing LSA in my other courses.

Input

As I worked on this project I continually sought after new sources. Books were often too dated for the Internet age discussing community only in terms of abstractions. Even in the last two weeks of the project serendipitous searching turned up two dissertations concerning signal-to-noise ratio on sites discussed in chapter 1 (Lawton; Malhotra et al., 1997). Besides these two, I discovered one dissertation early in the semester by corresponding with a user already on the Internet researcher network (Lackaff, 2005). I was quite happy with my introduction after reviewing these—many parallel concepts were independently deemed important and discussed.

Besides academia though, the Internet is a nearly infinite source of discussion on the topic of community. Despite the casual nature of these resources, they could not be discounted—my academic searching turned up little modern literature on discourse

communities since 2000. As the project drew to a close I used the data I had gleaned to contact individual users who had long ago left the primary community. Many of them had theories from which my project benefited from significantly.

Future Work

When I was feeling out the domain for this project initially, one of my professors with whom I met suggested that analysis was impossible without taking into account dynamical systems (Vallacher & Nowak, 1994). This forced me to reexamine the question of whether there are any demonstrable indicators that an online community is about to flourish, maintain, or fail. These are models such as those used to model chaos or muscular learning—useful when a given system is constantly changing as a function of itself or in some other complex fashion. For instance when a few of the primary content creators on Kuro5hin.org departed it may have resulted in the other core content creators leaving. This is logical and while this kind of data does surface in places like “Figure 18: Third Quartile post for top users,” it is hard to pin down the associations in any precise way. Future work might take this perspective into account.

Other methods which might be performed on the data include: Social Network Analysis; more sophisticated semantic comparison between comments and stories; or supervised learning including Bayesian methods or clustering. When polling the users from the site, one of my questions involved the formation of subgroups or cliques in the main portion of the site. It was my personal view going into the research that such subgroups were fairly rare on the main site (though common in the diaries). One of the more engaged users I interviewed suggested that they probably existed in some form

and wondered if any evidence would show that this influenced which stories were posted. Another agreed with minor reservations that they were not anywhere near exclusive cliques. “We kind of followed each other's comments like a rugby team follows the ball -- just ambling about in a big cluster.”

My LSA analysis was limited since there was no custom topic space containing modern Internet and cultural jargon. Many of the studies concerning Latent Semantic Analysis referenced other semantic methods using different decomposition techniques than LSA’s singular variable decomposition (Budiu, Royer, & Pirolli; Kaur & Hornof, 2005). Future work should generate a topic space from a source like Wikipedia and compare the comments and stories to confirm my tangentiality noise detection technique.

Supervised clustering involves manually marking content. This kind of analysis was clearly out of my own scope as a single programmer but it might be accomplished by a team of dedicated students. Bayesian methods are commonly used to sort spam and ham in email systems. This is somewhat similar to the crapflooding problem and perhaps slightly easier to detect. Clustering would be, as I understand it, more of a decision tree toward several automated classifications. If a message was just so long, included a certain number of topical words and was by a someone prominent in the community it would be in pile one.

Output

My goal is necessarily unattained. My search time was limited, insufficient to fully explore the large database I compiled. The “triangulation” of the various text processing methods will have to take place elsewhere. Readability shows some positive signs but the Latent Semantic Analysis test of tangentiality is relatively untested. Kuro5hin succeeded initially because of its fertile niche in high quality Internet reviewed articles and “failed” ultimately due partially to information overload and administrative neglect. I have cataloged a few other reasons using small scale surveys but the bulk of the users will go unquestioned. This failure caused the outflow of user base to other similar services which arose as the Internet matured and the site faltered. In some sense the site lives on but its niche is much tighter and its community tiny, barely still able to handle its malicious population. The problem is indeed difficult, all of the users departed for different reasons and determining the relative frequency of these is difficult to achieve passively.

Works Cited

Arendt, H. (1958). *The human condition.* University of Chicago Press Chicago.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven Conn.: Yale University Press.

Raluca Budiu, Christiaan Royer, Peter Pirolli (2007). Modeling Information Scent: A Comparison of LSA, PMI and GLSA Similarity Measures on Common Tests and Corpora. Proceedings of RIAO'07 – Large-Scale Semantic Access to Content (Text, Image, Video and Sound), Pittsburgh, PA.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods Instruments & Computers, 30*(2), 188-198.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new dale-chall readability formula*. Cambridge, Mass.: Brookline Books.

Chaskin, R. J. (2001). *Building community capacity*. New York: A. de Gruyter.

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization Science, 7*(2), 119-135.

Dick, P. K. (1968). *Do androids dream of electric sheep?* Oxford University Press.

Feenberg, A., & Bakardjieva, M. (2004). Virtual community: No 'killer implication'. *New Media Society, 6*(1), 37-43.

Feenberg, A., & Barney, D. D. (2004). *Community in the digital age: Philosophy and practice*. Lanham: Rowman & Littlefield.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). *The intelligent essay assessor: Applications to educational technology.* Retrieved 3/31/2007, 2007 from http://imej.wfu.edu/articles/1999/2/04/

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2-3), 285-307.

Geertz, C. (1993). *The interpretation of cultures.* Fontana Press.

Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*(1), 201-233.

Grubisich, T. (2006). *What are the lessons from dan gillmor's bayosphere?* http://www.ojr.org/ojr/stories/060129grubisich/

Gurstein, M. (2000). *Community informatics: Enabling communities with information and communications technologies*. Hershey, PA: Idea Group Pub.

Hague, B. N., & Loader, B. (1999). *Digital democracy: Discourse and decision making in the information age*. London; New York: Routledge.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243-1248.

Jones, Q., Ravid, G., & Rafaeli, S. (2001). Information overload and virtual public discourse boundaries. *Eighth IFIP Conference on Human-Computer Interaction,* , 9-13.

Jones, S. (1999). *Doing internet research: Critical issues and methods for examining the net*. Thousand Oaks, Calif.: Sage Publications.

Katz, J. E., & Rice, R. E. (2002). *Social consequences of internet use: Access, involvement, and interaction*. Cambridge, Mass.: MIT Press.

Kaur, I., & Hornof, A. J. (2005). A comparison of LSA, wordNet and PMI-IR for predicting user click behavior. *CHI '05: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Portland, Oregon, USA. 51-60. from http://doi.acm.org/10.1145/1054972.1054980

Lackaff, D. (2005). *Norm maintenance in online communities: Analysis of heterarchical moderation regimes.* La Trobe University; Melbourne, Australia. Retrieved 3/25/2007 from http://lackaff.net/node/25

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*. Cambridge; New York, NY, USA: Cambridge University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211-240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Landauer, T. K. (2006). *Handbook of latent semantic analysis*. Mahwah, N.J: Lawrence Erlbaum Associates.

Latham, R., & Sassen, S. (2005). *Digital formations: IT and new architectures in the global realm*. Princeton, N.J.: Princeton University Press.

Lawton, P. Capital and stratification within virtual community: A case study of metafilter.com. [Electronic version]. *Unpublished Dissertation,* Retrieved 4/2/2007,

Lee, Y., Chen, F., & Jiang, H. (2006). Lurking as participation: A community perspective on lurkers' identity and negotiability. *ICLS '06: Proceedings of the 7th International Conference on Learning Sciences,* Bloomington, Indiana. 404-410.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments & Computers, 28*(2), 203-208.

Malhotra, A., Gosain, S., & Hars, A. (1997). Evolution of a virtual community: Understanding design issues through a longitudinal study. *ICIS '97: Proceedings of the Eighteenth International Conference on Information Systems,* Atlanta, Georgia, United States. 59-74. Retrieved from http://drzaius.ics.uci.edu/meta/classes/informatics161_fall06/papers/17-1997.malhotraEvolutionVirtualCommunity.pdf

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, Mass.: MIT Press.

Mejias, U. A. (2005). Re-approaching nearness: Online communication and its place in praxis. [Electronic version]. *First Monday, 10*(3)

Negroponte, N. (1995). *Being digital* (1st ed.). New York: Knopf.

Nielsen, J. (2006). Participation inequality: Lurkers vs. contributors in internet communities (jakob nielsen's alertbox). Retrieved 10/21/2006, from http://www.useit.com/alertbox/participation_inequality.html

Powazek, D. M., & Safari Tech Books Online. (2002). *Design for community*. Indianapolis, Ind.: New Riders.

Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Putnam, R. D., Feldstein, L. M., & Cohen, D. (2003). *Better together: Restoring the American community*. New York: Simon & Schuster.

Raymond, E. S. (1999). *The cathedral & the bazaar: Musings on Linux and open source by an accidental revolutionary* (1st ed.). Cambridge, Mass.: O'Reilly.

Renninger, K. A., & Shumar, W. (2002). *Building virtual communities: Learning and change in cyberspace*. Cambridge, U.K.; New York: Cambridge University Press.

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Reading, Mass.: Addison-Wesley Pub. Co.

Schuler, D. (1994). Community networks: Building a new participatory medium. *Communications of the ACM, 37*(1), 38-51.

Sharkey, M. (2005). *Keepgoing.org : The big fish.* http://www.keepgoing.org/issue20_giant/the_big_fish.html

Sherman, R. C. (2003). *The Mind's Eye in Cyberspace: Online Perceptions of Self and Others*. Amsterdam: IOS Press.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications.* Wiley.

Takahashi, M., Fujimoto, M., & Yamasaki, N. (2003). The active lurker: Influence of an in-house online community on its outside environment. *GROUP '03: Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work,* Sanibel Island, Florida, USA. 1-10. from http://doi.acm.org/10.1145/958160.958162

Thurlow, C., Tomic, A., & Lengel, L. B.. (2004). *Computer mediated communication: Social interaction and the internet*. London; Thousand Oaks, Calif.: Sage.

Tönnies, F. *Gemeinschaft and gesellschaft. Encyclopedia Britannica.* http://www.britannica.com/eb/article-9036340/Gemeinschaft-and-Gesellschaft

Turkle, S. (1995). *Life on the screen: Identity in the age of the internet.* New York: Simon & Schuster, c1995.

Vallacher, R. R., & Nowak, A. (1994). *Dynamical systems in social psychology*. San Diego: Academic Press.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G. M., & Milios, E. E. (2005). Semantic similarity methods in wordNet and their application to information retrieval on the web. *WIDM '05: Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management,* Bremen, Germany. 10-16. from http://doi.acm.org/10.1145/1097047.1097051

Weber, M., & Parsons, T. (1947). *The theory of social and economic organization.* Free Press.

Whittaker, S., Terveen, L., Hill, W., & Cherny, L. (1998). The dynamics of mass interaction. 257.

Williams, F., Rice, R. E., & Rogers, E. M. (1988). *Research methods and the new media.* Collier Macmillan.

Ye, N. (2003). *The handbook of data mining*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Appendix I: The Dataset

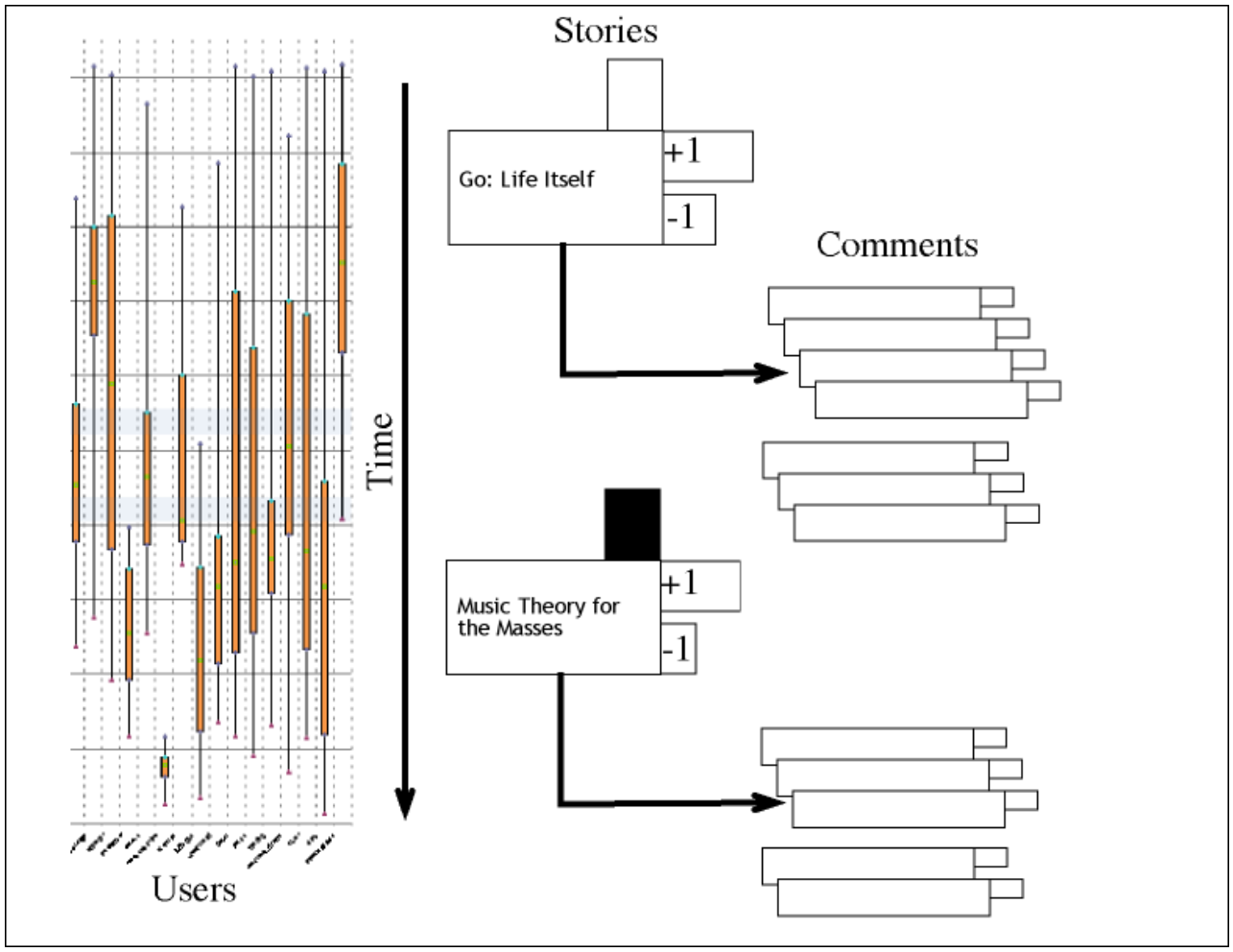

Figure 4: Data summary

Not pictured: Daily Web server hit count and (limited) visit count, monthly estimated overall population. I have rating data for each user, comment and story. The vertical bar represents the backdoor post algorithm which triggers if the story is in the queue for a long period.

Example Story

Paleotech 01: *http://www.kuro5hin.org/story/2002/4/6/155337/8442*

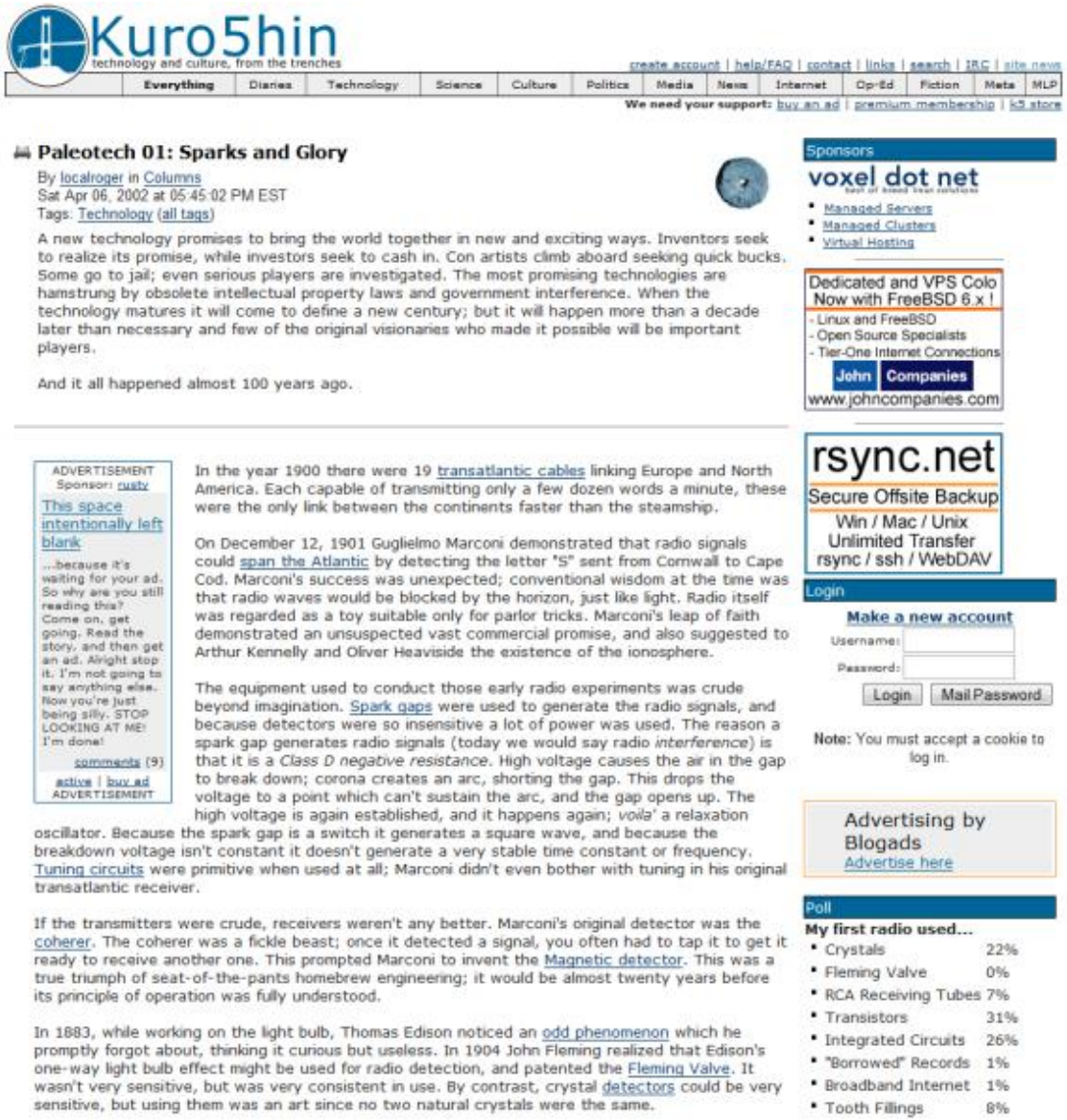

Figure 5: Example story

Processed Information:

Threshold:	95
Score:	**95**
Votes:	**156**
Comments:	**86**
Words:	**1764**
Kincaid scale	9.9
ARI scale	11.3
Flesch ease	60.2
Fog Index	13.3
Smog	11.9
Lix	46.4
Coleman	12.5

The Coleman method is the best consistent prediction of quality that I found (see Chapter 5, Results).

Example Ideal Comments

Processed Information

First Comment	
Max Possible Rating:	5
Rating:	**4**
Raters:	**1**
Words:	**1764**
Kincaid scale	10.6
ARI scale	11.8
Flesch ease	53.4
Fog Index	14.8
Smog	13.1
Lix	48.4
Coleman	13.8

Figure 6: Ideal comments

These comments above were rated 3 and 4 on the 1-5 scale. The comments below are on the 1-3 scale and are attached to a different more modern story near the end of the dataset.

Example Average Comments

My new theory of physics (2.37 / 8) (#5)
by b1t r0t on Tue Oct 31, 2006 at 05:04:02 PM EST

...is the "Finger Theory". If you want to see how it works, just pull my finger!

-- Indymedia: the fanfiction.net of journalism.

Our theory of the universe (2.09 / 11) (#6)
by United Fools on Tue Oct 31, 2006 at 05:11:52 PM EST

Things in general have no intelligence.

Can you prove to the world that you are not one of us ?

+1 fp (1.44 / 9) (#9)
by circletimessquare on Tue Oct 31, 2006 at 06:25:33 PM EST

a theory in search of a phenomenon to apply to

therefore, useless

string theory has no practical utility or implications or testable hypotheses, so its more like mythology for physics

I'm making a Low Budget HDV Filipino Horror Movie in NYC

- Well when by Comrade Wonderful, 11/08/2006 10:18:22 AM EST (**none / 1**)

First comment	
Max Possible Rating:	3
Rating:	**2.37**
Raters:	**8**
Words:	**1764**
Kincaid scale	0.5
ARI scale	0.2
Flesch ease	100.2
Fog Index	2
Smog	3
Lix	11.7
Coleman	8.2

Figure 7: Average comments

Below are general statistics I surreptitiously plumbed using a variety of indirect methods related to the central work on the project. I have the back end data in my database (it totals greater than 5mb).

Community's Sphere of Interest

General Section Headings

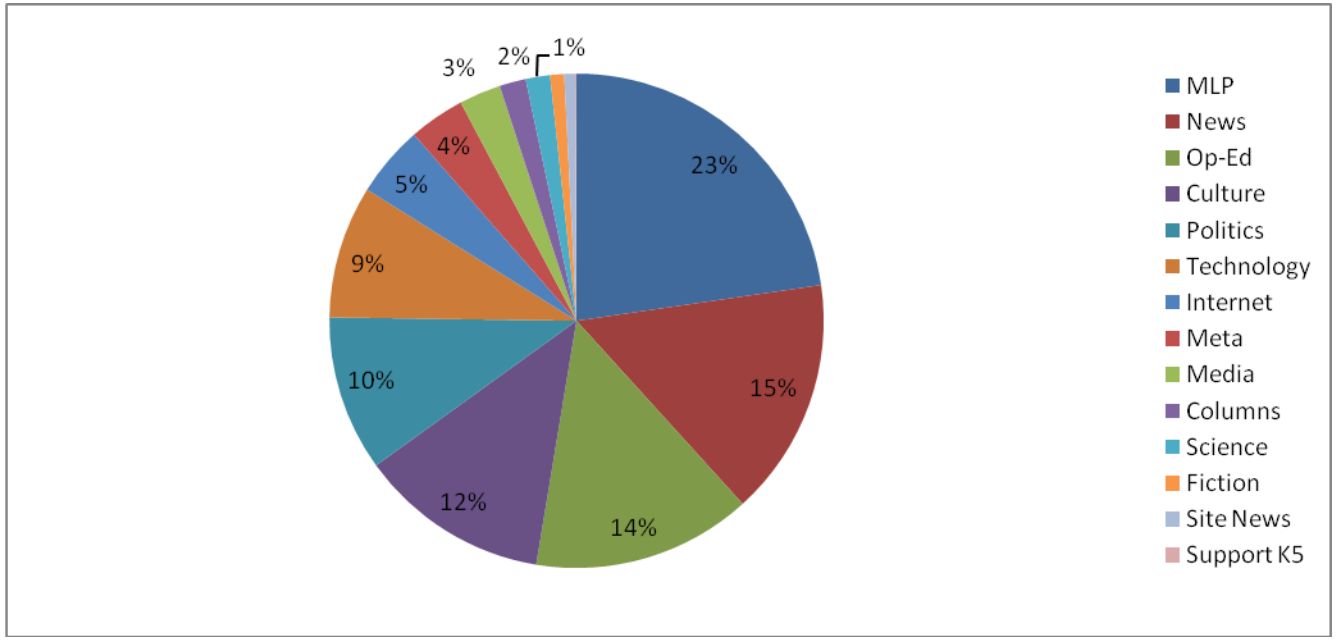

Figure 8: Section popularity

Sections work just like sections of a newspaper. They optionally divide the site into manageable pieces. Note that greater than 30% is related to Science and Technology. There are 6658 stories represented.

Sections

Sections (and Tags) are defined by the author of the article. If an article is improperly sectioned it is often voted down simply on that deficiency. Below are all of the sections (the primary classification of an article). They are listed in descending order of words written for those sections.

1. **Culture** stories have dominated the site for some time and cover everything from recipes to instructions on how to play the strategy game Go.
2. **Op-Ed** articles are opinion pieces (or rants) which often cover topics from other categories but are especially speculatory.
3. **Technology** articles include HOWTOs on programming, reviews of gadgets, historical articles on technology or speculative pieces on the future of some technology.
4. Articles in the **Politics** section discuss political issues in the United States and abroad.
5. **MLP or Mindless Link Proliferation** is a category which has dominated similar online story oriented discussion so it is mildly discouraged. Despite this, probably due to the brevity of the posts and the ease of publication it is the largest section by story count though not by word count (fifth largest).
6. Stories in **News** are also fairly brief, cover a current event somewhere online or in the world and often are intended to spark discussion.
7. **Internet** contains stories concerning the world wide web.
8. **Science** is for stories discussing scientific research.
9. **Fiction** is a relatively new section which accommodates fiction from users.
10. **Media** articles typically critique various forms of mass media.
11. **Columns** provide a place for people to do a continuing piece on some topic.
12. **Meta** articles discuss Kuro5hin, the Internet, or community abstractly on K5.
13. **Site News** announces changes to the structure or function of the site.
14. **Support K5** was a special topic created during the fund drive of 2002.

General Category Tags

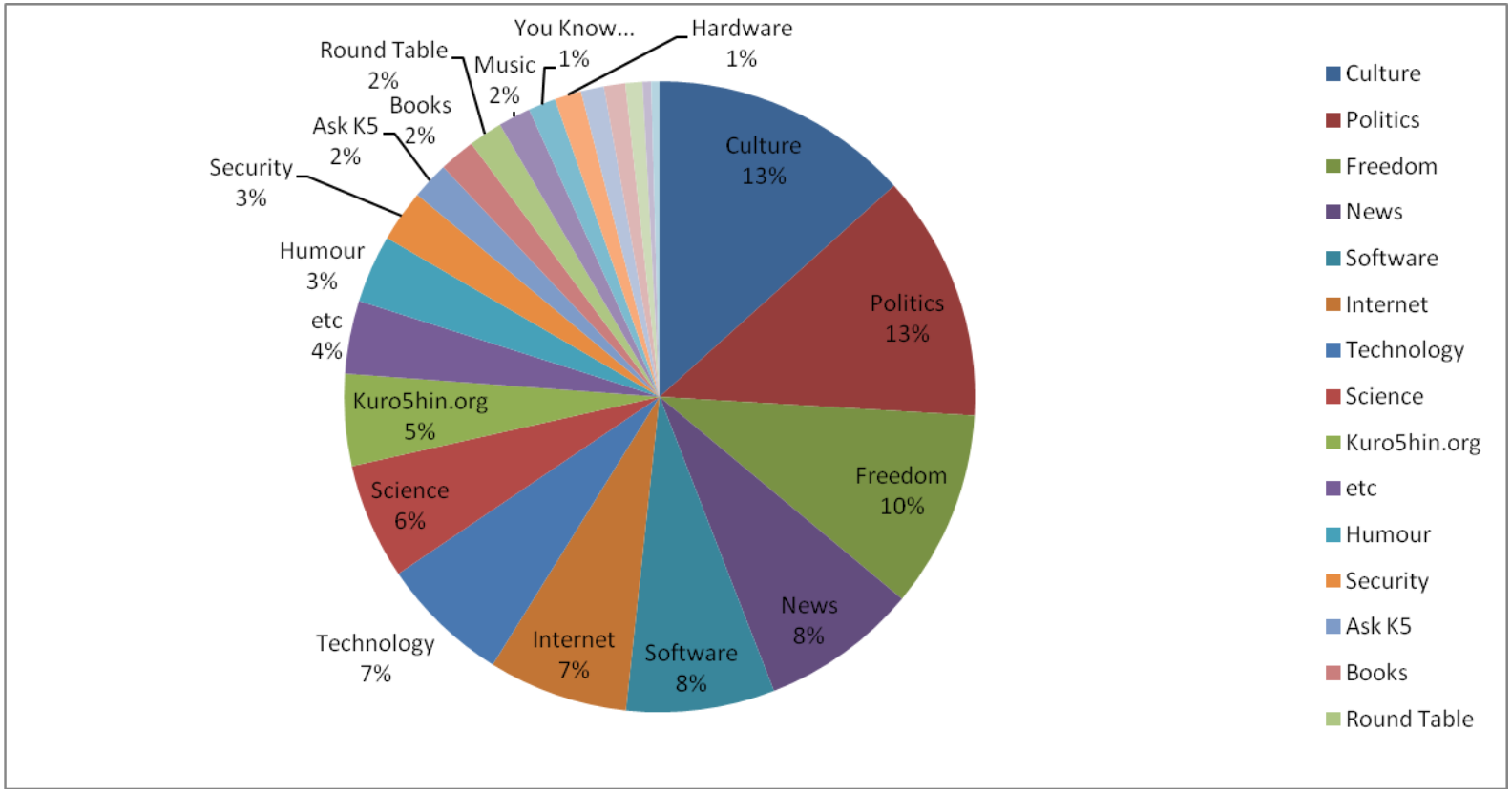

Figure 9: Specific tag popularity

Tags indicate a more flexible subordinate categorization of a story. They used to be called “categories.” Note again, Software, Internet, Science and Technology are a large portion but Politics, Freedom and Culture are just as large. The site is hardly strictly technical topics. This may lead to difficulties getting a good dataset for grading. There are 6658 total stories represented.

Appendix II: Kuro5hin.org statistics.

Change over Time

Users over Time (Logarithmic Scale)

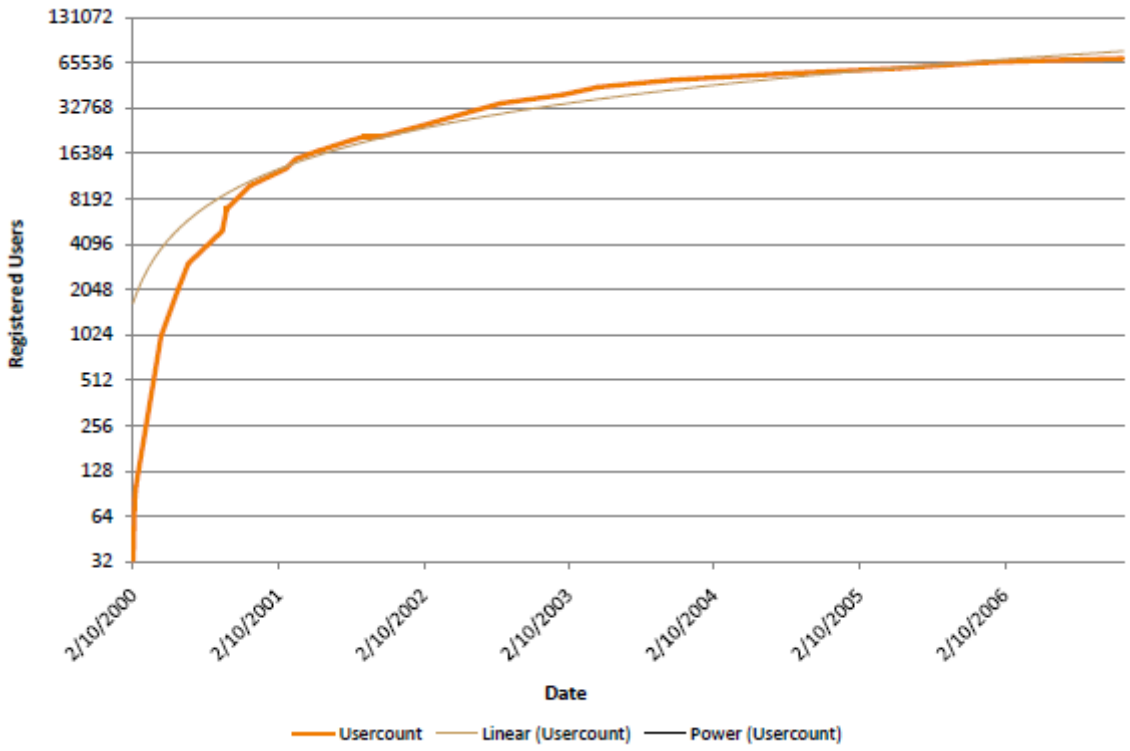

Figure 10: Registered users over time

This data was acquired with 32 samples of a user of a given user number's first posting. It was checked by examining several of the adjacent user's dates of first posting. Note that it is not exactly linear—It was increasing faster between 2002 until 2003.

New accounts per Day

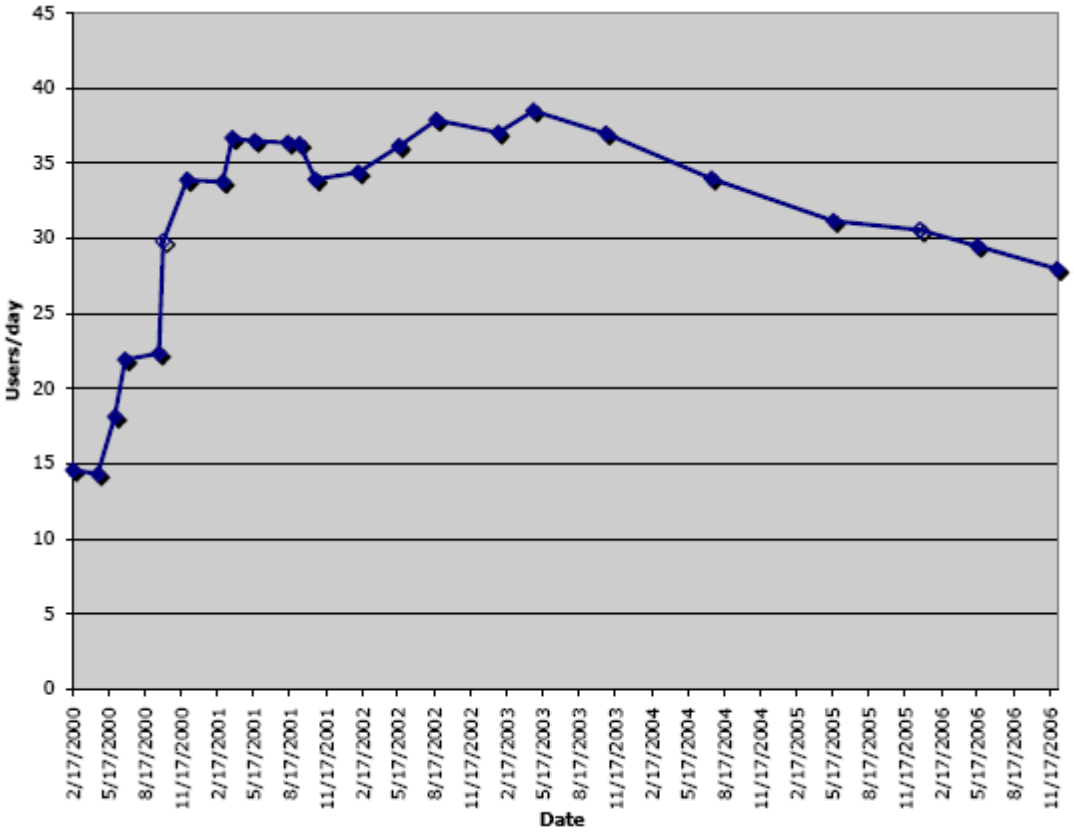

Figure 11: New accounts per day

This is the instantaneous estimate for the rate of new accounts created per day. It has been steadily decreasing since 5/17/2003 and is now at a low comparable to the point at the middle of the site's second year.

Stories and Comments per Day (Logarithmic Scale)

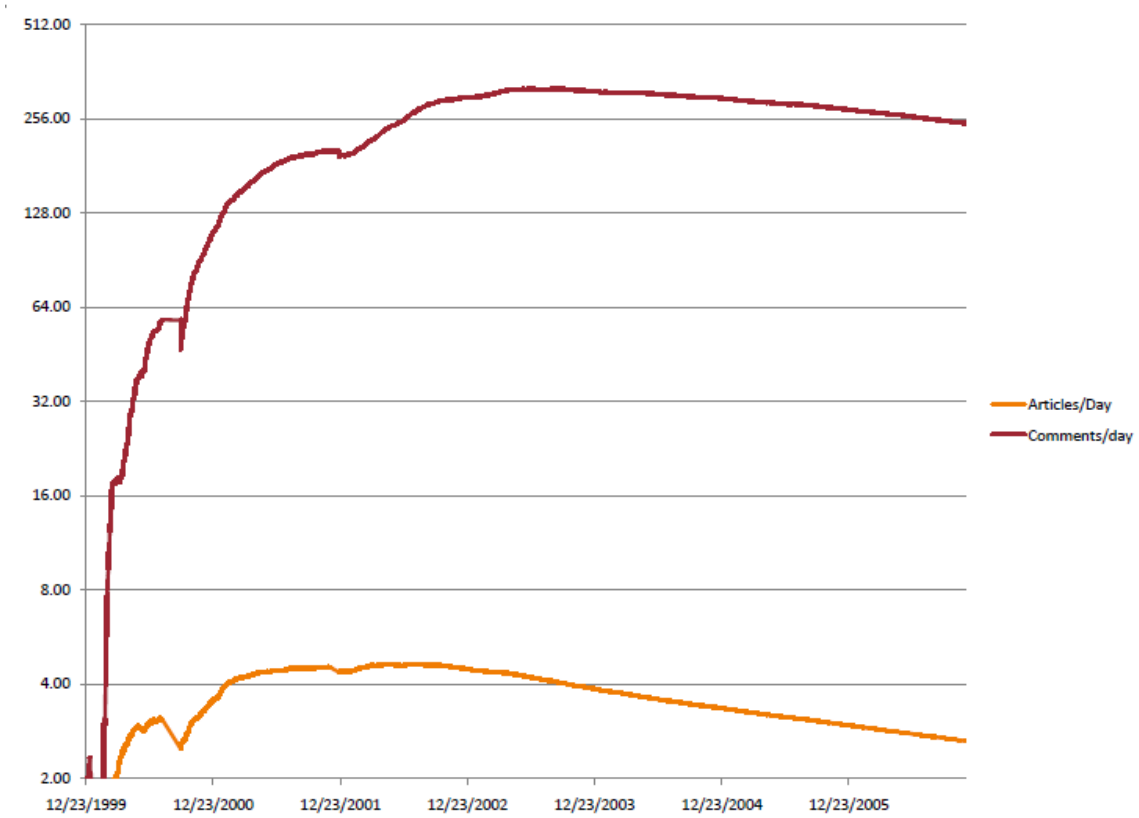

Figure 12: Stories and Comments per day

Here we have the estimated instantaneous changes in the posting of Stories and Comments per day. The story posting reached a local maximum of 4.5 on 11-22-2001 (perhaps as a result of the September 11 discussion), took a dip to 4.3 during the following spring (mostly due to twenty days of site maintenance), and finally reached pinnacle of 4.64 on 7-28-2002 by 12-2006 it was at 2.6/day (9-2000 levels). Even today the average is still 3.9 stories/day. New comment volume peaked at 319/day, hold an all-time average of 200/day but now sit at 247/day, still fairly high (6-2002 levels).

General Activity

Word Count of Stories Over Time

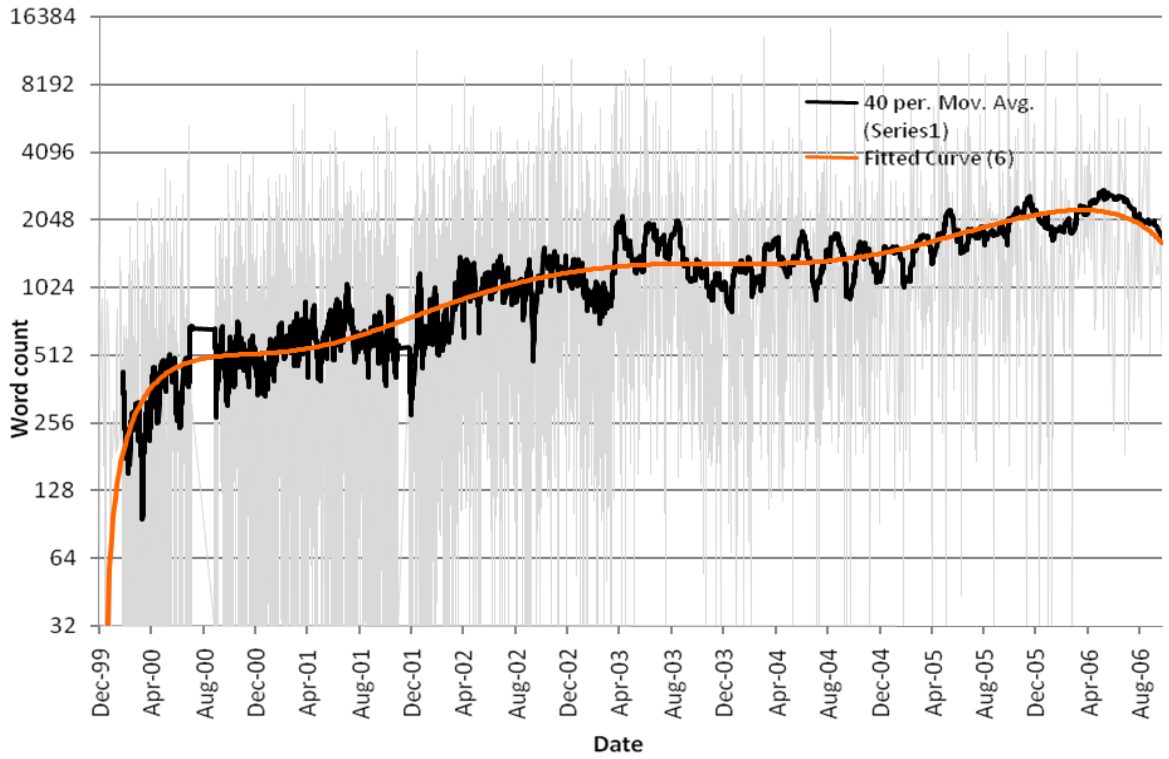

Figure 13: Word count in stories

The story length had been steadily increasing (sometimes indicating a well thought out story) until recently. This demonstrates the changes in demographics of the site over time.

Comments on a Story Over Time

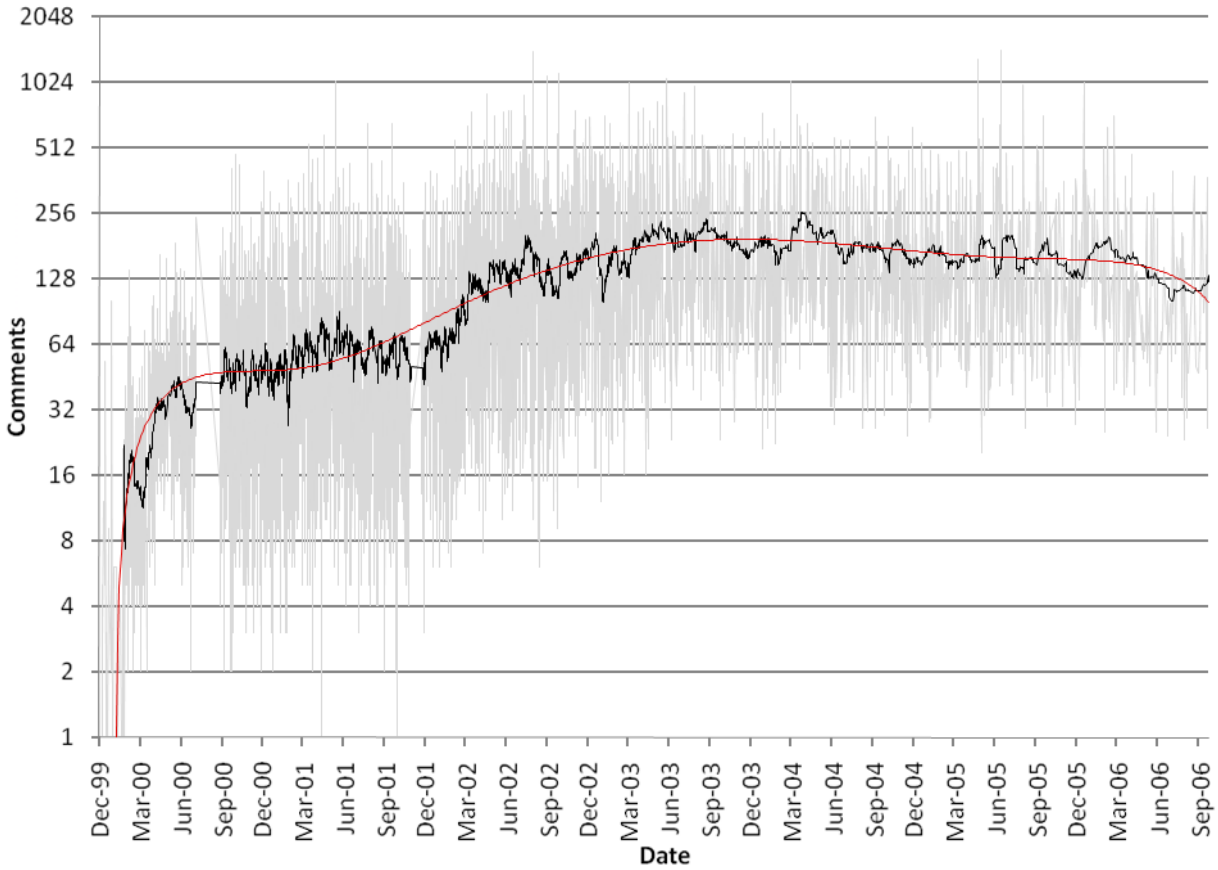

Figure 14: Comments on a story over time

Comments on a story has been flatter since 2001 but may be starting to dip. This chart is different from comment volume which trends like the story count. Black line is a forty day moving average.

Individual User's Story Contributions

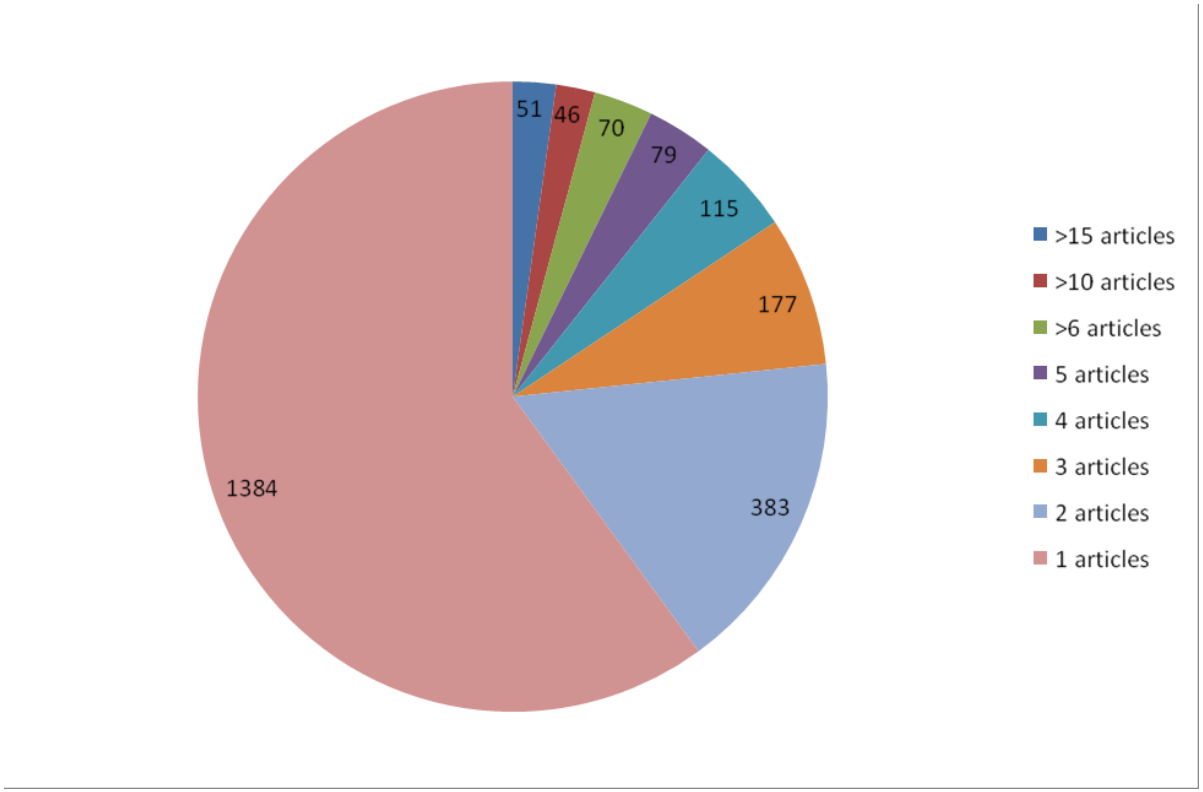

Figure 15: Participation scale

Behind 69,000 users there are 2305 main contributors. The top 51 users produced 1664 (24.992%) of the articles posted to the site. See "Chapter 6: Discussion Concerning the Failure of Community" concerning active participation.

Superfigure

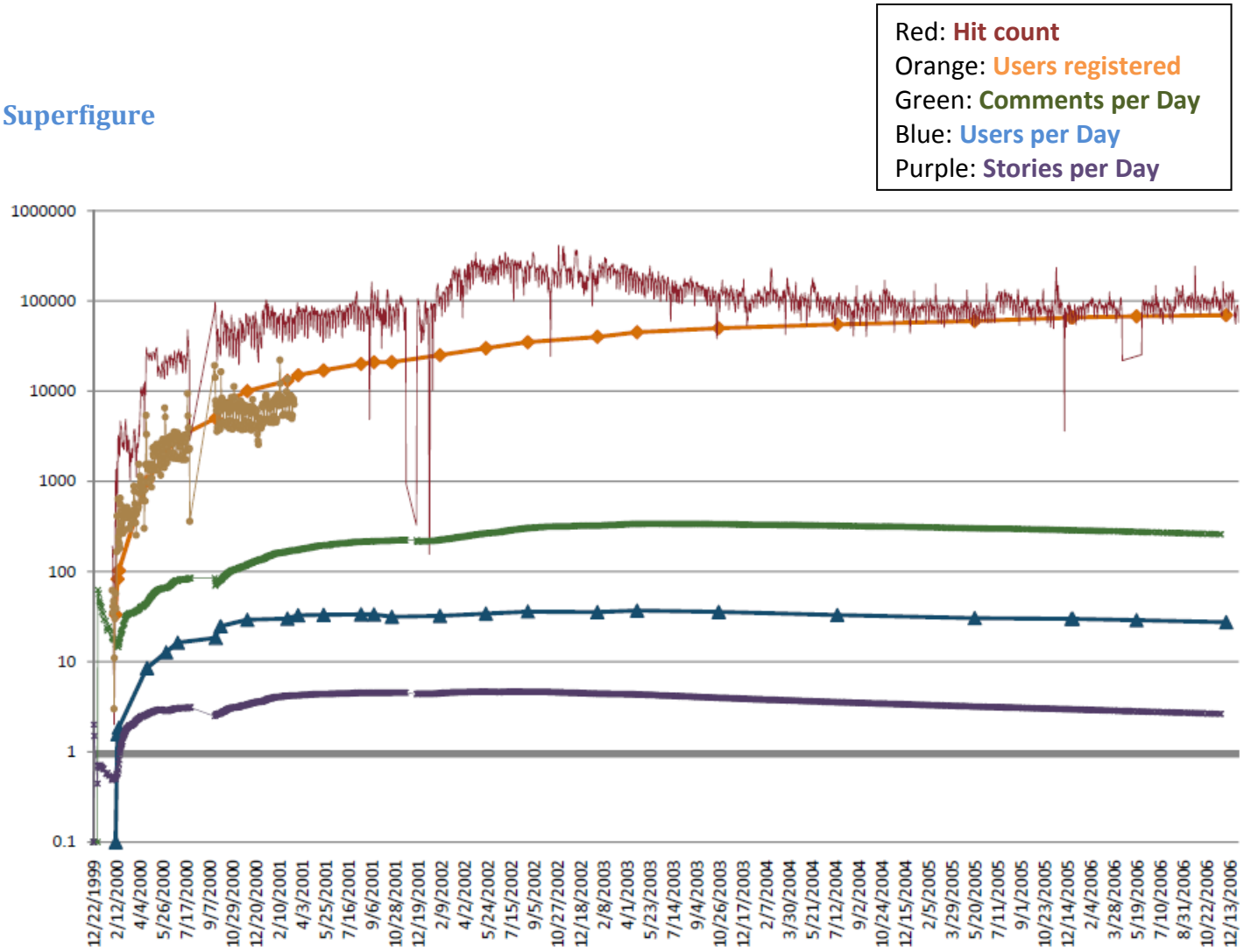

Figure 16: Broad dataset figure on logarithmic scale

This figure shows much of the interesting population data all at once. Please disregard the "Max of" and "Average of" on the labels, these are not directly relevant to the chart. Some trends are deemphasized due to the requisite logarithmic vertical axis. Note that the Hit count for the website has been relatively constant since 2003.

The Corpses of Pseudonyms

User Landscape and lifecycle (excerpt)

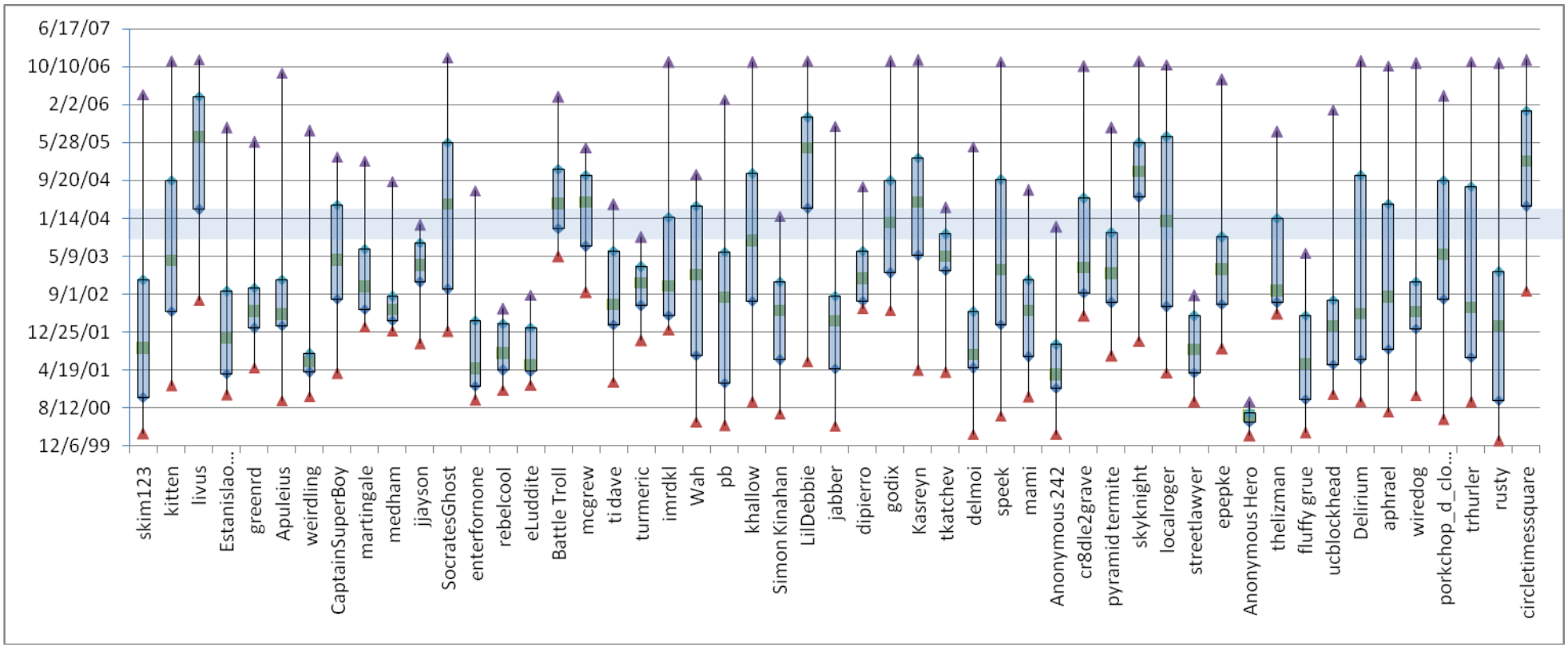

Figure 17: Activity date boxplots for top fifty users

Minimum, Maximum, First (25%), Median and Third (75%) Quartile post dates for the top fifty posters. A social network analysis would probably find these useful once participants are grouped off. The light blue backdrop shows the period with the greatest drop in participation. This is a novel technique as far as this project was researched.

Users decrease in posting

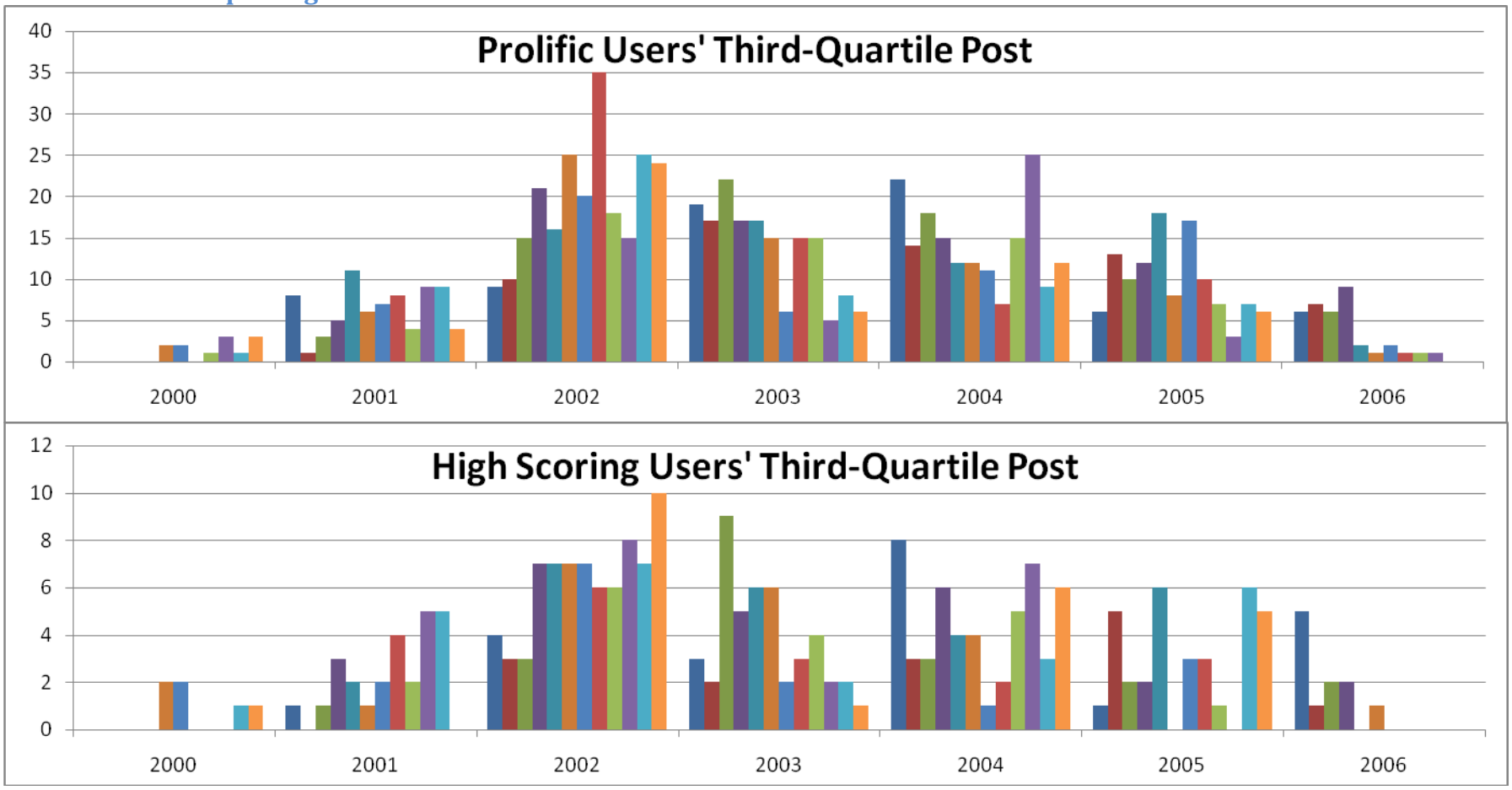

Figure 18: Third Quartile post for top users

Top: 800 users with the most comments (>150) date of final post. These uses accounted for 351,092 of ~600,000 posts (58%). Bottom: 250 highest scored users (based on comment count, comment rating, story count and story grade level) date of final post. The high scoring users accounted for 209543 posts, or 36% of the total posts. There are notable peaks and valleys which might demarcate certain low or high points of the community.

Total Departure of Users

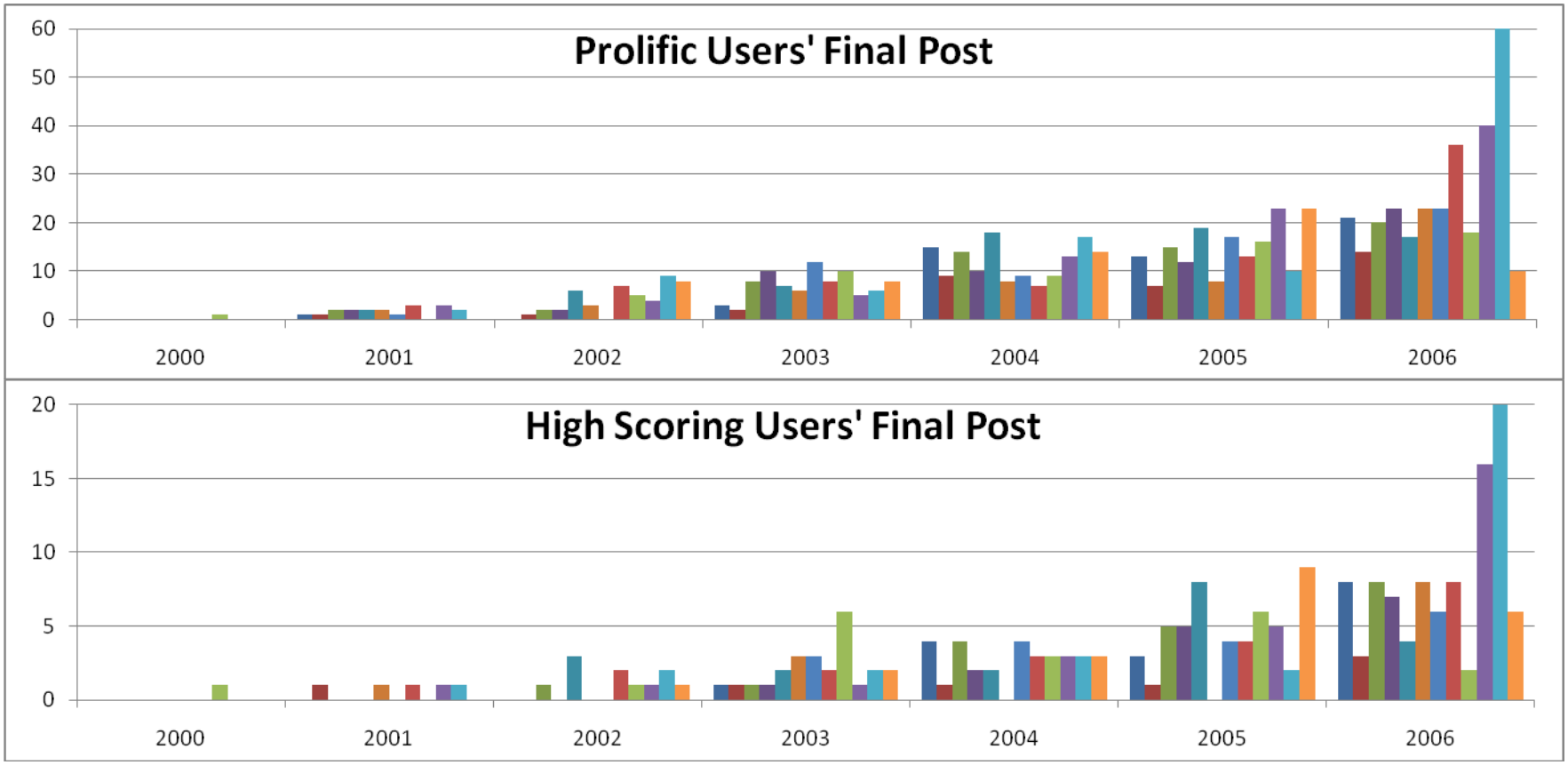

Figure 19: Final post for top users

Top: 800 users with the most comments (>150) date of final post. 150 of these users (19%) had posted in the last three months of the sample date range. Bottom: 250 highest scored users (based on comment count, comment rating, story count and story grade level) date of final post. 70 of these users (28%) were active in the last three months of the sample date range. This chart is more constant, showing the steady departure of the participants.