Is Wikipedia Growing a Longer Tail?

Shyong (Tony) K. Lam John Riedl

GroupLens Research
Department of Computer Science and Engineering
University of Minnesota
{lam,riedl}@cs.umn.edu

ABSTRACT

Wikipedia has millions of articles, many of which receive little attention. One group of Wikipedians believes these obscure entries should be removed because they are uninteresting and neglected; these are the *deletionists*. Other Wikipedians disagree, arguing that this long tail of articles is precisely Wikipedia's advantage over other encyclopedias; these are the *inclusionists*. This paper looks at two overarching questions on the debate between deletionists and inclusionists: (1) What are the implications to the long tail of the evolving standards for article birth and death? (2) How is viewership affected by the decreasing notability of articles in the long tail? The answers to five detailed research questions that are inspired by these overarching questions should help better frame this debate and provide insight into how Wikipedia is evolving.

Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software—*Information networks*; H.5.3 [**Information Systems**]: Group and Organization Interfaces—*computer-supported collaborative work*

General Terms

Human Factors, Measurement

Keywords

Wikipedia, long tail, collaboration, evolution

1. INTRODUCTION

Wikipedia has emerged as one of the world's most popular destinations on the web. With millions of articles on a staggering variety of topics, Wikipedia has successfully established itself as a useful compendium of general knowledge, and is read in excess of 150 million times per day [15]. Some might even consider its plethora of information to be an addictive time sink[1].

[1]http://xkcd.com/214/

GROUP'09, May 10–13, 2009, Sanibel Island, Florida, USA.

1.1 Wikipedia and Related Work

The fascination with Wikipedia extends beyond its readers. Researchers from numerous academic fields have also taken notice of Wikipedia, recognizing it as a ecosystem consisting of many interesting processes that are ripe for study.

Community Collaboration. Wikipedia's core idea is that anyone can edit just about anything on the site, and that *all* of its content comes from the people who use it. At first thought, this sounds like a recipe for disaster – how can a group of ordinary people write a good encyclopedia? Why would someone volunteer to do this? What about people who do not know what they are doing? How can we know whether people are writing good articles?

Despite these challenges, Wikipedia's reliance on its user community seems to be working, and researchers have worked to learn more about how and why. A study by Kittur, et al. found that as much as 50% of the early work on Wikipedia was done by a tiny group of "elite" contributors making up less than 5% of its editor population [10]. However, the study finds that in recent times, the balance has shifted toward a larger number of infrequent contributors, with the work done by the elites declining to less than 30%. Bryant, et al. studied several individuals' motivations for "becoming Wikipedian," learning about difficulties and successes they experienced along the way [4]. Viegas, et al. introduced a visualization tool to help understand patterns in how people's edits to an article have shaped and reshaped the article over time [16].

Conflict and Vandalism. Of course, letting anyone edit any article also has downsides. There can be disagreement among those who are working on an article, leading to conflicts and arguments about the article's content. Kittur, et al. have studied ways to identify and visualize conflict, finding that it has been on the rise as Wikipedia grows [11]. Vuong, et al. developed models to detect the presence of controversy by looking at how people add and delete words when editing an article [17].

Also, some users are malicious and will vandalize Wikipedia articles by deleting content, adding nonsensical content, or injecting misinformation. Viegas, et al. used their visualization tool to identify and study the edit patterns commonly used by vandals and the response they receive from the rest of the community. Their results suggest that Wikipedians are fast at repairing the damage vandals cause, with more than half of mass content deletions being addressed within three minutes [16]. Priedhorsky, et al. discovered that the visible impact of vandalism in Wikipedia to its readers is small but rising rapidly [14].

Governance. Dealing with issues such as conflict and vandalism lead to the need for governance. There must be policies about what types of behavior are acceptable or unacceptable, processes to guide how conflicts are resolved, and guidelines regarding article content and style. Wikipedia is largely self-governing, with

many decisions being made by the users themselves. A number of studies have focused on this self-governance, seeking to learn about things such as about how consensus forms and evolves [6, 12], or how policies are used to guide collaboration [3].

Content Quality. Finally, the content itself in Wikipedia is also of much interest. Some see it as a vast source of general knowledge and wonder if the semi-structured content can be used in interesting ways. For instance, Milne, et al. developed ways to automatically generate domain-specific thesauri from Wikipedia articles. They found that the generated thesauri offered good coverage and more contemporary language usage than expert-created ones [13].

However, because Wikipedia's content is not necessarily written by experts, much debate exists whether it is complete, accurate, and high-quality. Giles compared the quality and accuracy of Wikipedia and *Encyclopedia Britannica* and found that to the statistical limits of the study, the number and distribution of errors in the two encyclopedias was comparable [7]. Wilkinson and Huberman find that one distinguishing factor of high-quality Wikipedia articles is a larger number of editors [18], which is perhaps counterintuitive given the old adage: "too many cooks spoil the broth."

1.2 The Long Tail

In our work, we look at an area related to the wide breadth of article content on Wikipedia, but that has ties to the other processes described above as well: community collaboration, conflicts, and policies. Specifically, we explore *the long tail* of Wikipedia articles. We look at what the long article tail is and how it is evolving as people create more articles. We also study how the long tail is being affected by the policies that govern what topics are suitable for articles, and the ongoing conflicts among people who disagree about these policies.

Before we formally state our research questions, we describe the phenomenon of the long tail and how we apply it in the context of Wikipedia. The Long Tail is a term introduced by Chris Anderson, the editor-in-chief of *Wired*, that refers to a business strategy that received much attention in recent years [2]. The long tail gets its name from the natural long-tailed and heavy-tailed distributions appearing in consumption rates of many types of products such as books, songs, or movies. In these distributions, a small number of popular "hits" dominate, while a large number of items that individually have little consumption form the so-called "long tail." Many of these distributions are power laws.

The long tail strategy is to bolster a business's performance by finding ways to offer items from the long tail at little to no cost to the business. The theory is that while each long tail item only gets purchased a few times, their aggregate sales can significantly increase revenue. Traditionally, this strategy has not been viable for physical stores because selling everything in the tail requires too much floor space. Thus, retail stores must carefully choose what items they stock, and they naturally gravitate toward high-volume "hits" from the head of the distribution.

However, the long tail strategy is usable by web-based retailers because physical storefronts are not required. Warehouse space is cheaper than retail space, so it becomes possible to stock and offer more items. Furthermore, in some domains such as digital music, the cost of storing items is negligible since the items consume little to no physical space.

The long tail has also been used to describe phenomena in non-commerce domains such as blogs, social networks, and tagging. Here, the long tail often refers to the natural long-tailed distributions found in these domains rather than to a business strategy. For instance, the influence of blogs is described as a long tail by Agarwal, et al in their work to identify influential bloggers [1]. Golder and Huberman found that in collaborative tagging systems, distributions of tag usage can be modeled using a stochastic urn process that naturally yields long-tailed distributions [8].

In the context of Wikipedia, we apply the long tail to its collection of encyclopedia articles and the viewership that each article receives. As a point of reference, consider that the 2008 Britannica Encyclopedia Suite contains 65,000 articles[2], which is well less than 5% of the millions of articles in the English Wikipedia. Using a Wikipedia web log dataset that we describe in section 2, during the last three months of 2007 the top 65,000 Wikipedia articles ranked by visits comprise less than 60% of all visits to Wikipedia articles. So, if we consider the remainder of Wikipedia articles to be the long tail, it makes up over 40% of Wikipedia traffic, which is about 60 million article views per day as of late 2008 [15]!

Other researchers have explored different manifestations of the long tail in their studies of Wikipedia. Kittur, et al's analysis suggested the emergence of a long tail of user participation in which much work is being done by a large group of people, each of whom only does a small amount of work [10]. Wu, et al. found that the use of infoboxes, a specific type of structured data found in Wikipedia articles, follows a long tail distribution, and proposed ways to improve automated information extraction methods in the presence of such a skewed distribution [19]. The present paper is the first research to look deeply at the questions of the long tail in article readership, combined with an exploration of how issues of article mortality influence the evolution of the long tail.

1.3 Research Questions

In the rest of the paper, we first discuss the datasets we use for our analysis, and then address the following five research questions, in one section each.

RQ *Long Tail Visits*: To what extent do Wikipedia viewers look at articles in the tail?

RQ *Wikipedia Growth*: How have article birth and mortality rates changed over time?

RQ *Topic Notability*: As time passes, are the articles that survive in Wikipedia increasingly on obscure topics?

RQ *Deletion Reasons*: What are the reasons given for deleting articles? How do these reasons relate to the long tail?

RQ *Article Life Span*: When in the life of an article is it most likely to be deleted?

The present research is different from many other projects that study group dynamics in that it is a study of a single distinctive community. Because of the distinctive – some say unique – properties of Wikipedia, it is not obvious how to extrapolate these results to other communities. We argue that Wikipedia is such an important group activity, with millions of visitors every *day*, that research that helps understand how and why it works is independently interesting, even if it is not obviously generalizable. We further speculate that the successes of Wikipedia, if deeply understood, can lead to the design of computer support for other groups that can share some of those successes. However, the scope of the present paper is limited to deepening our understanding of Wikipedia.

2. DATA

Most of our analyses are performed using the information sources that are described below.

English Wikipedia data dump files. These are several datasets that were made available on the Wikipedia database download site[3] at various times between 2006 and 2008. Each dump contains a

[2]http://tinyurl.com/britannica-size

[3]http://download.wikimedia.org/enwiki/

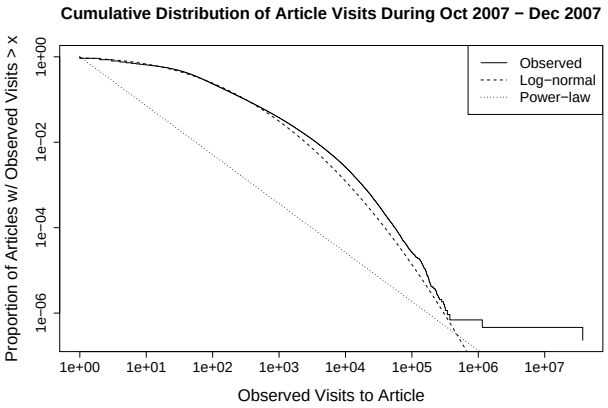

Figure 1: Complementary cumulative distribution function of Wikipedia article visits on log-log scales. Visits between October 1, 2007 and December 31, 2007 are counted. The dashed and dotted lines are maximum-likelihood estimate fits to the log-normal and power-law distributions.

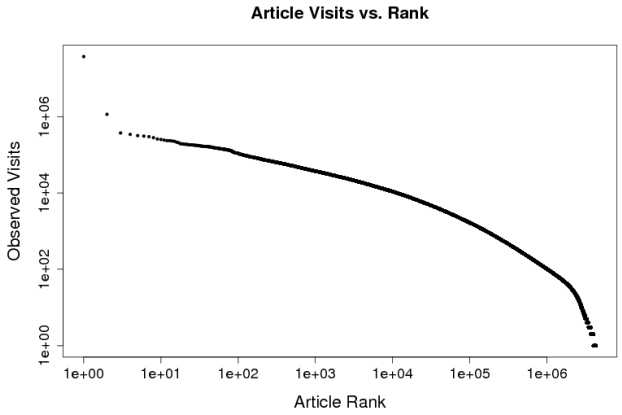

Figure 2: Rank-frequency plot of Wikipedia article visits on log-log scales. The top two most visited pages, which appear to be outliers, are the "Main_Page" and "Wiki" articles.

snapshot of all articles that existed on Wikipedia when the dump was created. The data provides information about every article revision (i.e., time of creation, author, and edit comments).

In this paper, all of our analysis is confined to the *Main* namespace. We make this distinction because we are interested specifically in the encyclopedic content available on Wikipedia, which is stored in the *Main* namespace. Other Wikipedia namespaces are not considered in our analyses. These namespaces include ones such as: *Talk*, which contains meta-discussions about articles; *User*, which contains personal information about Wikipedia users; and *Wikipedia*, which contains information specific to Wikipedia itself (e.g., help pages, community standards, content guidelines).

English Wikipedia event log. The Wikipedia download site also provides a log of special events that have occurred on Wikipedia. These events include administrative meta-actions such as blocking abusive users, renaming articles, or granting users with special flags or privileges. Of particular interest to our analysis, this log also contains information about article deletions, including the given reason for deletion.

We note that by deletion, we mean the case where an administrator has removed the article *and* all of its revision history from public view. We do not consider cases where a user simply erases all the text from an article, since the removed text is still readily available by browsing the article's revision history.

Sample of Wikipedia web logs. The Wikimedia Foundation has graciously supplied us with an anonymized feed of the web access log for their web servers. The feed contains the URL and timestamp of every 10th HTTP request. This log allows us to accurately estimate how many people are reading each Wikipedia article. Note that because of this sampling, all reported viewership figures are approximately a factor of ten below their actual values.

3. LONG TAIL VISITS

First, we will look at the overall distribution of visits across Wikipedia articles to get a sense of what Wikipedia's long article tail looks like. For this analysis, we used a web log sample from October 1, 2007 through December 31, 2007.

To visualize this distribution, we present it as a complementary cumulative distribution function (CCDF), which is shown in figure 1. The value of the CCDF is equal to one minus the value of cumulative distribution function. When plotted on a log-log scale, this representation of the data allows us to see whether the distribution is a power law by looking at whether the CCDF is a straight line. Furthermore, this representation is more robust against biases and noise in the data than the probability density function or rank-frequency plots [5, 9].

Figure 1 also shows maximum-likelihood estimate fits to the power law and log-normal distributions. The CCDF shows a distinct curve, and is a much better fit to a log-normal distribution than to a power law. This result is in apparent conflict with a Wikipedia article that examines the traffic of the top 1,000 most visited articles and concludes that beyond the top few articles, the distribution looks like a power law[4]. The reason for the disagreement is that often power law and log-normal distributions appear the same when looking at only a few orders of magnitude on a log-log scale. When the Wikipedia data are extended to the full set of articles, rather than just the top 1,000, the function is clearly non-linear (figure 2), and a power law fit can be ruled out.

The observation that Wikipedia traffic is not a power law raises interesting questions about Wikipedia's evolution. The conditions lend themselves naturally to a long tail power law scenario: practically unlimited storage for articles, low barrier to entry, and efficient digital distribution. Yet, the empirical results suggest the distribution is much closer to log-normal, which manifests itself as a truncated power-law distribution or a "drooping tail" in which there is a deficiency of low-readership articles. One possibility is that the natural distribution would be a power law, but that other factors such as efforts to deter creation of low-value articles have "truncated" the distribution such that it has become log-normal (for a discussion of the evidence on this point, see section 6).

Overall, in answer to **RQ *Long Tail Visits***, Wikipedia traffic does show a long-tailed distribution, especially over the first thousand articles, but do not follow the classic power law over the entire six orders of magnitude of article popularity. A log-normal distribution is a better description of the data. Some authors argue that only power law distributions should be called long-tailed, while others argue that log-normal distributions should share the name. We won't get caught up in argument about terminology here, but will note that excluding the 65,000 most popular articles from Wikipedia – 65,000 is the number of articles in the entire Encyclopedia Britannica – still leaves 60 million article views a day for the rest of Wikipedia. So, Wikipedia traffic is substantially increased by the long tail phenomenon.

[4]http://en.wikipedia.org/w/index.php?oldid=222154521

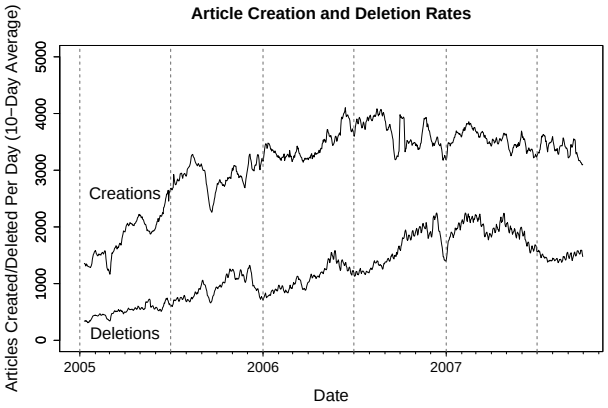

Figure 3: Daily rates of surviving article creation, and article deletion. Rates are smoothed using a ten-day moving average.

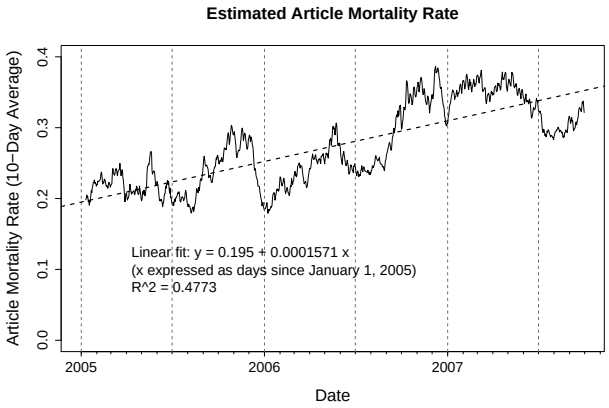

Figure 4: Estimated article mortality rate, smoothed using a ten-day moving average.

4. WIKIPEDIA GROWTH

Before we delve deeper into Wikipedia's evolution over the years, we first look at the big picture – how has Wikipedia grown? What broad patterns have there been in the creation and deletion of articles over time?

4.1 Data Challenges

A major issue with the article dumps is that they do not contain any information about articles that were deleted prior to the time that the snapshot was made. The deletions are captured in the event log, but information about the articles that were deleted is not present. This leads to several challenges in performing analyses of article deletion behavior, since there is limited information about the vast majority of articles that have been deleted.

Data about deletions occurring prior to December 2004 are unavailable, so we only focus on the period between December 2004 and December 2007 for which we have both creation and deletion data. Also, robots and users of semi-automated editing tools[5] occasionally perform tasks that introduce substantial noise to our data. These tasks include one-off projects such as creating articles about politicians or animal species by copying information en masse from outside sources. Because the these tasks undergo advance review by administrators and fellow community members[6], the created articles are typically not scrutinized, and are rarely subject to deletion. In order to focus on how Wikipedia's long article tail is affected by the actions of human users, we exclude articles created by known automated processes in our analyses that involve deletion of articles.

4.2 Article Birth and Death Rates

Figure 3 shows the birth rate of surviving articles[7] and the death (deletion) rate of articles, expressed as a ten-day moving average to smooth out noise. We present the birth rate of surviving articles because the true article birth rate is not known due to the lack of data on deleted articles. Specifically, we do not have data that tells us when deleted articles were originally created, so we cannot determine the total number of articles created during a given interval. However, the available data do yield some interesting patterns.

[5]http://en.wikipedia.org/wiki/WP:AWB

[6]http://en.wikipedia.org/wiki/WP:BOT

[7]Our article counts differ from Wikipedia's published statistics due to differences in the definition of an article. We only exclude bot-created articles, whereas Wikipedia includes bot-created articles, but excludes some shorter entries that they consider "non-articles."

Note that the death rate tends to follow the surviving birth rate, rising and falling mostly in lockstep. This correlation suggests that if an article death occurs, it tends to be near the time that the article was created. Later, in section 7, we use other datasets to enable careful measurement of survival rates and allow us to validate this conjecture. For now though, we will simply assume that article deletions occur near the time of article creation. Making this assumption allows us to estimate the total article birth rate during some period by summing the death rate and the surviving birth rate.

With estimates of article birth and death rates, we can compute an estimated article mortality rate and see how it changes over time. Figure 4 shows this relationship. While there is much fluctuation in the mortality rate, there is a modest upward trend in mortality, suggesting that new articles are being increasingly subject to deletion as Wikipedia grows and evolves. This is consistent with results presented in Kittur, et al. that show that Wikipedians are spending an increasing amount of their efforts on indirect work – enforcing policy, dealing with vandalism, and so on [11].

We also find that there is noticeable movement in this mortality rate that correlates to actions taken by the Wikimedia Foundation or its members. In particular, we found the following two instances.

First, in December 2005, the Wikimedia Foundation made the decision to restrict article creation to users who have a Wikipedia account[8]. This was done in response to a high-profile vandalism incident involving an article about former *USA Today* editor John Seigenthaler, Sr. In May 2005, somebody created a hoax article about Seigenthaler that linked him to the John F. Kennedy and Robert F. Kennedy assassinations. The article was left untouched for several months before Seigenthaler learned about it from a colleague. After working with the Wikimedia Foundation to have the article removed, Seigenthaler published an op-ed article in *USA Today* describing the incident and criticizing Wikipedia[9].

Figures 3 and 4 show that during December 2005, the article mortality rate fell by roughly 30%, while the surviving article birth rate remained unaffected. It is plausible that the new restriction dissuaded would-be vandals or pranksters from creating questionable articles such as the hoax about Seigenthaler, and therefore reduced the number of articles that required deletion. However, the reprieve was only temporary, as the mortality rate began rising again soon afterward. Perhaps the barrier of account creation was insufficient as a long-term deterrent to undesirable articles.

[8]http://en.wikipedia.org/w/index.php?oldid=136017357

[9]http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm

Second, in August 2006, Jimmy Wales, co-founder of Wikipedia, gave a keynote talk at the Wikimania conference during which he urged Wikipedia contributors to focus on article quality rather than article quantity. Wales' keynote received coverage by the mainstream media, with articles appearing in the *New York Times*[10], *Wired*[11], and other outlets. Looking back at figures 3 and 4, we see that in August 2006, Wikipedia's article birth rate decelerated and the death rate accelerated, leading to a noticeably elevated article mortality rate that remained high for about ten months. It seems that Wikipedians agreed with Wales and raised the bar for what constituted an acceptable Wikipedia article.

We stress that we have no solid evidence that these actions were directly responsible for the changes observed in Wikipedia activity. These are interesting correlations that suggest that external factors may have a profound effect on the evolution of Wikipedia.

In answer to **RQ *Wikipedia Growth***, we can say that while the number of articles in Wikipedia is growing, their mortality rate is also slowly increasing over time. Of course, from these data we cannot say whether the Wikipedians are applying tougher criteria to new articles, or whether the newly created articles are less appropriate for Wikipedia. We shall return to that question in the next section.

5. TOPIC NOTABILITY

5.1 Deletionism and Inclusionism

The observation that over one-quarter of Wikipedia articles are ultimately deleted leads us to look at a long-running conflict that has been taking place on Wikipedia for years. The constant influx of thousands of articles per day is a source of concern for some Wikipedians who believe that many articles are about topics that are too obscure and that are not interesting enough to warrant a Wikipedia article. These people see such articles as diluting the overall value and credibility of Wikipedia. Others believe that the ever-growing set of articles is a good thing since it allows more opportunities for people to participate, and emphasizes Wikipedia's strengths as a digital resource that has no practical limit on size. These two philosophies have been labelled as *deletionism* and *inclusionism*, respectively, and the results of their influences on Wikipedia and its long tail will be the primary focus of the remainder of this paper.

We now look more closely at how Wikipedia and its long article tail have evolved over time. How do articles that were created years ago compare to more recently created articles? How true are deletionist concerns that Wikipedia's newer articles are increasingly about obscure topics? What did Jimmy Wales see that triggered his call for focusing on quality rather than quantity?

5.2 Data Challenge: Notability

To approach these questions, we need a way to measure the relative obscurity or popularity of an article. For this, we first turn to Wikipedia's standards regarding this issue. Wikipedians use a basic criterion called *notability* to decide whether a particular topic is worthy of an article. There are a wide range of opinions on the definition of notability and how much it should be taken into account when deciding whether an article belongs in Wikipedia. Much debate between inclusionists and deletionists has taken place on Wikipedia regarding notability, and the notability guidelines are often invoked when discussing whether to keep or delete an article that has come under scrutiny.

[10]http://www.nytimes.com/2006/08/07/technology/07wiki.html

[11]http://www.wired.com/science/discoveries/news/2006/08/71535

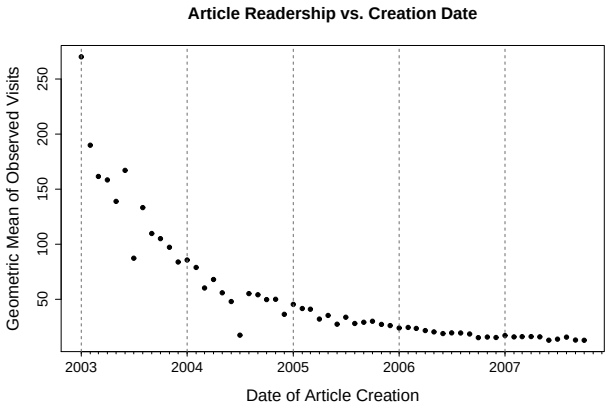

Figure 5: Geometric mean of the readership of articles plotted by month of article creation.

To pass Wikipedia's general notability guideline[12] as of late 2008, an article's topic must have "received significant coverage in reliable sources that are independent of the subject". Wikipedians have also established additional domain-specific notability guidelines for things such as books, films, and numbers[13]. These guidelines, while well-articulated, are often imprecise and open to interpretation (e.g., what exactly constitutes "significant coverage"?).

Thus, in this paper, we do not propose a way to directly operationalize notability. Instead, we will use metrics that measure *popularity*, which is a related notion that may correlate well with notability in practice. While it is true that popularity is not exactly the same as notability, and that the metrics we use are unreliable in individual cases, we believe that our metrics are a good proxy for notability if taken in aggregate.

5.3 Readership

The first metric we will consider is readership. We measure this by counting the number of visits to each article as given in the Wikipedia web log sample, again using the interval October 1, 2007 through December 31, 2007. Articles that are read more frequently are presumed to be about things that are more well-known and interesting to Wikipedia readers, so this metric estimates how popular or obscure an article's subject is.

Figure 5 shows the average readership of Wikipedia articles as a function of when the articles were created. There is a striking downward trend indicating that newer articles are being viewed far less frequently on average than older ones. This suggests that newer articles tend to be about topics that draw less interest from Wikipedia readers, and are thus more likely to be in the long tail.

One confound here is that newer articles are disadvantaged because they have had less time to integrate themselves into the link structure of Wikipedia, and thus, have fewer backlinks (i.e., other Wikipedia pages linking to them). This deficiency of backlinks may result in newer articles receiving less traffic since users browsing Wikipedia encounter fewer links to new articles than to old ones. In turn, one might surmise that traffic to new articles will start low and accumulate over time as more links are created. To investigate, we control for the backlink effect by repeating our analysis but grouping sets of articles that have similar numbers of backlinks. Figure 6 shows that a similar downward trend still holds, although

[12]http://en.wikipedia.org/wiki/WP:N

[13]This particular guideline was, in part, prompted by a deletion debate over an article about the number 3.14, a common approximation of the mathematical constant Pi.

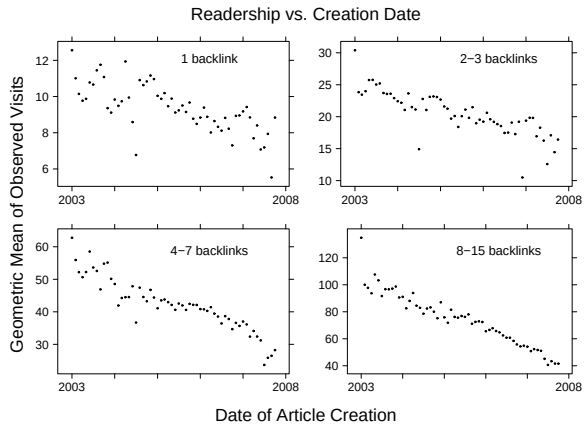

Figure 6: Geometric mean of the readership of articles plotted by month of article creation, grouped by articles with similar numbers of backlinks. Plots of articles with 1 backlink, 2-3 backlinks, 4-7 backlinks, and 8-15 backlinks are shown.

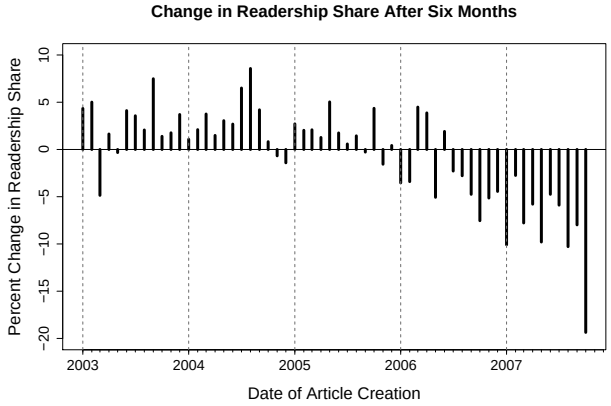

Figure 7: Relative change in total readership share between October-December 2007 and July-September 2008, plotted by month of article creation. Only articles created between January 2003 and October 2007 are considered.

the drop over time is smaller, and now appears more linear for each group. Apparently there is an important effect of number of backlinks in explaining article traffic, but it alone does not fully explain the readership differences between older articles and newer articles.

Additionally, we did an analysis to quantify how article readership changes over time. (Articles may gain readership because of increases in backlinks within Wikipedia, because of links from the Web as a whole, because of improving position in search engines, etc.) We hypothesize an asymptotic effect exists where articles gain readership for some time before approaching a stable state that represents its "true" popularity. Thus, over time, newer articles should gain readership while older articles remain stable (effectively losing readership relative to the whole population).

To test our hypothesis, we compared the readership figures described above with figures from July 1, 2008 through September 30, 2008 to see what had changed after six months. Figure 7 shows the relative change in article readership share between the two data sets, again as a function of article creation date. We see that contrary to expectation, older articles increased their readership share at the expense of newer articles, which actually lost readership share as they aged! One possible explanation for this is that newer

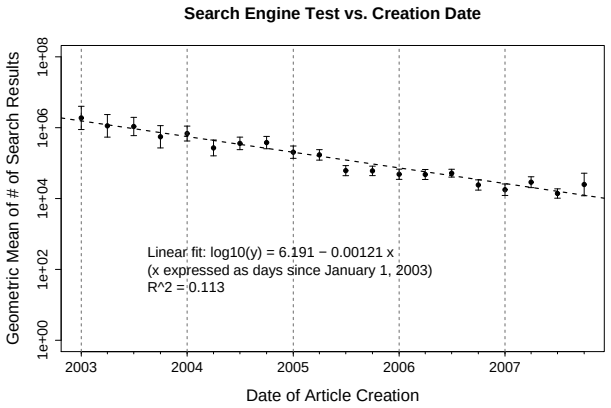

Figure 8: Geometric mean of results of Search Engine Test plotted by month of article creation. Geometric standard error bars and a best-fit line computed from unaggregated log-transformed data are plotted as well.

articles tend to be about things that naturally have initial bursts of interest, such as current events and new movies or video games. Overall interest in such topics then declines over a period of time before reaching some stable state. Further research is needed to confirm or refute this explanation, but in any case, our results do not show evidence that measuring popularity using the readership metric is biased against newer articles.

5.4 The Search Engine Test

A second metric that we use to approximate notability is the search engine test. This is also known on Wikipedia as the Google Test[14]. This metric is defined as the number of results that a search engine returns when queried for web pages about a particular topic. The search engine test provides an estimate of popularity that has the advantage of being mostly independent of Wikipedia. (The presence of Wikipedia and sites that copy its content inflate the values, but their effect is probably small compared to the size of the web.)

However, Wikipedia's article about the search engine test gives several caveats in using it to establish the popularity or notability of a topic, and states that the test's result alone should not be considered to be authoritative. One major issue described is that "search engines do not disambiguate, and tend to match partial searches." The Wikipedia discussion provides a simple example: the Renaissance painting *Madonna of the Rocks*. Depending on how a search engine query is formulated, there might be many search results about the pop singer *Madonna*, which would inappropriately make it appear as if this painting was much more popular than other well-known Renaissance paintings.

We attempt to control for this problem by restricting our analysis to articles that have single-word titles. While there is still opportunity for ambiguity (e.g., *jaguar* could refer to an animal, car manufacturer, or football player), we believe it reduces the effects of the problem sufficiently for our purposes. Also, using single-word titles eliminates the challenge of formulating queries for multi-word titles (i.e., word order or use of quotes), as well as confounds arising from differences in the distribution of the number of search results for multi-word searches versus that for single-word searches.

We chose a random sample of 5,758 articles with single-word titles and issued basic queries against the Yahoo! search engine using

[14]http://en.wikipedia.org/wiki/WP:Google_test

Table 1: Classes of Deletion Reasons.

Class	Deletion Reasons
Inappropriate Content	Patent nonsense; vandalism; attack pages; blatant advertising; copyright infringement
No Content/Context	Insufficient context to identify subject of article; insufficient substantive content
Notability/Significance	Failure to assert importance or significance; non-notable subject
PROD/AFD/VFD	Proposed deletion; articles for deletion; votes for deletion
Wiki Maintenance	Redirect to a non-existent page; technical deletion (used for renaming or moving articles, merging article histories, and other maintenance-related tasks)
Other	Creator requests deletion; creation of previously deleted material; all other policies
Unknown	No recognized key words or key phrases

their API[15]. We were unable to test all single-word title articles in a reasonable amount of time due to limitations imposed by Yahoo!'s usage policy. Figure 8 shows the relationship between the mean number of search engine results and the Wikipedia article creation date. We see a downward trend similar to the one shown previously for article readership, thus reinforcing the support for the results obtained using the readership metric: newer articles tend to be more concentrated in the long tail and are effectively lengthening it.

The readership data and the search engine test both provide the same answer to **RQ *Topic Notability***. Apparently new articles that are added to Wikipedia *are* increasingly obscure, and are thus likely to be less notable. We do note that these data alone do not resolve the debate between inclusionists and deletionists. After all, the long tail of not-so-popular articles is responsible for a substantial number of Wikipedia page views. However, the data might provide a principled way to reason about the cost versus value of adding articles to Wikipedia. For instance, this question could be put on an economic footing by valuing article readership in dollars, and by estimating the monthly cost of the resources required to maintain each article. Ones attracting insufficient interest to justify their cost would be deleted. (Economic motivations are not the only way to select articles that belong in an encyclopedia; this is just one possible way to frame the debate.)

6. DELETION REASONS

The deletionists have likely seen evidence of these notability trends, and argue that Wikipedia is increasingly becoming a haven for irrelevant material that should have failed the test for notability. Some deletionists are working hard to seek out articles they feel are not notable, and to remove them from Wikipedia. In this section we study their success at this task, looking at the frequency of deletes, the reasons for deletes, and the changes across time in these characteristics. We are particularly interested in the effect these changes are having on the evolution of the long tail in Wikipedia.

Wikipedians have established several different processes for deleting articles[16].

Criteria for Speedy Deletion. This is the most lightweight process for deletion. There are several dozen reasons for which an article can be deleted without requiring a discussion. Among these include vandalism, advertising, or insufficient content. This process is intended to be used for uncontroversial deletions.

Proposed Deletion (PROD). This process is used when somebody believes that an article should be deleted, but for a reason not covered by the Criteria for Speedy Deletion. If no one objects to the proposed deletion, then the article is deleted. If there is an objection, the issue is escalated to the Articles for Deletion process.

[15]http://developer.yahoo.com/search/

[16]http://en.wikipedia.org/wiki/WP:DP

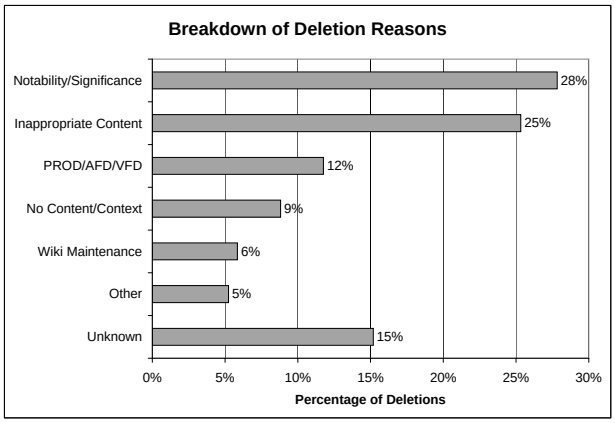

Figure 9: Overall frequency of classes of reasons given for Wikipedia article deletions. PROD/AFD/VFD denotes deletions occurring as a result of the Proposed Deletion or Articles for Deletion processes.

Articles for Deletion (AFD). In this process, interested members of the community examine the article under scrutiny and discuss what should happen to it. Discussions last at least five days, after which time an administrator reviews the debate and takes appropriate action. This process was previously also known as "Votes for Deletion" (VFD), but was renamed because the goal is to make decisions based on community discourse rather than majority vote.

To analyze why articles are deleted, we use the event log dataset, which includes the comment left by the deleter for each deletion event. The comment is intended to convey the reason that the article was deleted. By analyzing these comments, we can gain insight into why over one-quarter of all created articles are deleted.

We scanned deletion comments for key words or phrases that refer to Wikipedia's article deletion policies. For example, the deletion comment for an article that was deleted via the Proposed Deletion process typically contains a link to the Wikipedia policy page that describes the process, *WP:PROD*. Thus, we can identify such deletions by looking for a *WP:PROD* link. We also looked for other textual indicators of this process, such as "proded", "prodded", and "proposed deletion". We created similar lists of key words for identifying other reasons for deletion. Approximately 85% of the deletions studied could be categorized in this way.

In total, we looked at 1,567,543 deletion comments for deletions occurring between December 2004 and March 2008. Using our approach, we classified deletions into seven broad classes, which are summarized in table 1.

Figure 9 shows the overall frequency that each of these classes of deletion reasons was observed in the deletion comments. Only 12% of deletions go through the more heavyweight processses (Pro-

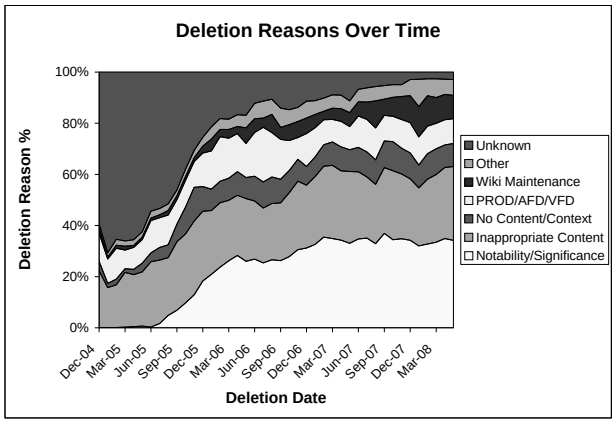

Figure 10: Frequency of classes of reasons given for Wikipedia article deletions by month. PROD/AFD/VFD denotes deletions occurring as a result of the Proposed Deletion or Articles for Deletion processes.

posed Deletion, Articles for Deletion, or Votes for Deletion). A large majority of deletions are considered uncontroversial and are covered by the Criteria for Speedy Deletion. We see that the most frequently-cited reasons for deleting an article are notability-related, making up over a quarter of all deletions. Next, deletions due to inappropriate content (25%) or insufficient content (9%) together make up just over a third of article deletions. Wiki Maintenance and Other are both around 5% each. Finally, 15% of deletions, labelled "Unknown" in the figure, could not be categorized using simple keyword analysis.

Figure 10 shows the relative frequency of deletion reasons across time. We see two noteworthy trends here.

First, the proportion of unknown deletion reasons is declining, which means an increasing proportion of deletions are accompanied by recognized citations to Wikipedia policy. This trend is consistent with the findings in Beschastnikh, et al. that show an temporal increase in policy citations on Wikipedia discussion pages [3].

Second, the proportion of deletions due to reasons classified as Notability/Significance has increased over time. As we saw in section 5 (figures 5 and 6), there has been a lengthening of the long tail as article creators push the boundaries for what is considered notable. Our observations here suggest that some in the community are pushing back, actively scrutinizing articles and deleting those that are deemed not notable enough.

One thing that we cannot tell from studying deletion reasons, however, is whether interpretation and application of the notability guidelines has been consistent. Are the articles that are being deleted *actually* less notable than the articles that survive? One way to approach to this question is to apply our notability proxy metrics to articles that have been deleted due to lack of notability.

The readership metric is difficult to use here, because as we will see in section 7, the lifetime of an article before it is deleted is usually too short to gather meaningful data. We can easily apply the search engine test though, as it does not depend on Wikipedia-specific data. On a random sample of 959 articles with single-word titles that were deleted due to lack of notability, the geometric mean of the number of search results is 6,832. This is below the average number of search result for surviving articles in Wikipedia, which, according to figure 8, is well over 10,000, even for the most recently created articles. The comparison suggests that the deletion decisions being made regarding notability are generally consistent with the search engine test.

However, we note that if the downward trajectory seen in figure 8 continues at its historic pace, then articles created in mid-2008 will have an average number of search results of around 6,000, which is comparable to that of articles that have been deleted in the past for lack of notability! Over the long term, the declining notability of new articles will lead to one of two possible outcomes. An inclusionist might hope that notability standards will become less stringent. On the other hand, a deletionist might hope that notability criteria will remain stable, and that a higher percentage of newly created articles will be deleted.

These data provide a mixed answer to **RQ *Deletion Reasons***. Overall, the "lack of notability" reason has dramatically increased in usage between 2005 and the present. However, its increase has been very slow since early 2006, and nonexistent since early 2007. The distribution of reasons given for article deletions appears to have reached a steady state. Also, deletion decisions seem to be consistent with the search engine test for topic notability.

7. ARTICLE LIFE SPAN

Finally, we explore the life span of Wikipedia articles and look at *when* articles get deleted during their lifetimes. How quickly does the community scrutinize new articles and make decisions about them? Was the Seigenthaler incident the norm or the exception? Are deletionists trimming the long tail, or is it here to stay?

7.1 Data Challenge: Article Creation Dates

Recall that our data is deficient in that we do not know the creation date of most deleted articles. We overcome this limitation of our datasets in three ways:

Direct Data Analysis. First, we directly use the Wikipedia dumps to obtain what information we can about article life span. By combining an older snapshot of Wikipedia articles with a newer event log, we can see which of the older articles have been deleted after the time that the snapshot was taken. This gives us life span information about long-lived articles, but only provides limited and flawed information about short-lived articles.

To illustrate this issue, suppose that we want to learn about articles with a life span of less than 2 days. The only articles we could examine are those that were created during the 2 days immediately preceding the time that the article snapshot was taken. If an article was created before this interval and was deleted within 2 days, then it would not have appeared in the snapshot. To make matters worse, there is an additional confound: articles that were created *and deleted* during the window of interest would also be absent from the snapshot and missed by the analysis, which leads to an undercount of articles with a life span of less than 2 days.

Inference-Based Analysis. To help augment our knowledge about very short-lived articles, we also use an inference-based approach using the article snapshots and event logs. Consider a snapshot taken at time t containing the set of all articles A existing at time t, and an article deletion event for some article a that occurs at time $u = t + 1$ day. If $a \in A$, then we know the creation date of a, and can use the log analysis approach previously described.

On the other hand, suppose $a \notin A$. Then we do not know the creation date of a. However, we do know that a must have been created after time t, since by definition A contains all articles that existed at time t. We also know that a must have been created before time u because an article cannot be deleted before it is created. Therefore, despite not knowing its exact creation time, we infer that a's life span is less than 1 day.

Applying this logic to all articles deleted in the first n hours after an article snapshot allows us to count how often articles were created and deleted during that n hour interval. This provides a basis

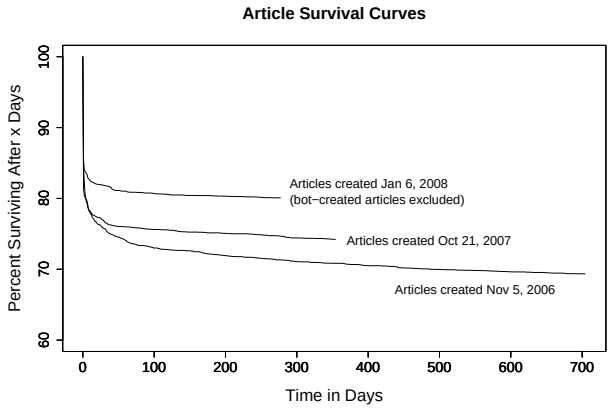

Figure 11: Survival curves of Wikipedia articles created during three 24-hour spans. The first-day death rates are estimated as described in section 7, while all remaining data is observed.

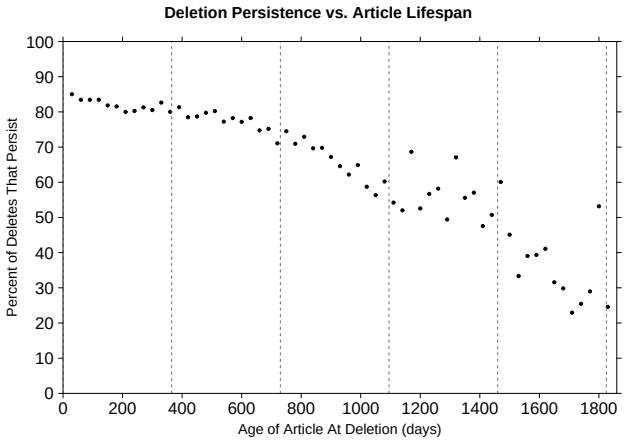

Figure 12: Persistence of deletions of Wikipedia articles, plotted by age of article at deletion.

for making estimates about articles that have very short life spans. However, this approach can be used just once for each article snapshot, and only provides information about articles over a small slice of time. It is, therefore, subject to the same issues that plague small sample sizes – high variability and questionable precision.

Near Real-Time Observation. To help solidify our data about short-lived articles, we turned to the Wikipedia API[17], which can be queried for information about article creations that occurred during a given interval. This data is subject to the same shortcomings as the data dumps: articles that have been deleted do not appear in article creation listings. However, the adverse effects of missing data can be greatly reduced by issuing API queries often, thus capturing article creation events in approximately real-time. For our analyses, we collected article creations every five minutes over a two week interval in September 2008. We then used an event log from October 2008 to determine whether the created articles had been deleted, and if so, when.

7.2 Life Span Results

Combining all the amassed information, we found that most articles have either a very short life or a very long life. If an article is deleted, then the deletion usually occurs very early in the article's life, quite often within the first few days. Recall that in section 4, we conjectured that if an article is to be deleted, then the deletion will occur near the time that the article was created. Here, we will present data that supports this supposition.

Analysis of our inference and real-time observation data shows that for any given 24-hour period, about 61% of deletions during the period are targeted at articles that were created during that period. This allows us to generate estimated survival curves with our article snapshots. The bottom-most line plotted in figure 11 shows our estimated survival curve for articles created during the last 24 hours before the November 2006 article snapshot. The first-day deaths are estimated, but the remainder of the curve is actual data. Interestingly, over 20% of articles survive less than a day, and about 25% survive less than two weeks. Beyond that, just another 5% of articles are deleted over the following two years.

Figure 11 also shows survival curves generated similarly from October 2007 and January 2008 snapshots. The shapes of the curves are similar, although they "flatten out" at different percentage levels. This reflects the volatile mortality rates shown previously in figure 4. In all three survival curves, we see that a large majority of deaths occur during the first few days of an article's life. Wikipedians make inclusion and deletion judgments about articles very quickly, and it is uncommon for the community to return to articles later and delete them.

We also examined the question of whether deletions are "persistent" – that is, if an article is deleted, does it stay deleted, or does someone create the article again later? To measure deletion persistence, we compared our 2006 and 2008 article snapshots and looked at which of the articles existing in 2006 had been deleted in the interval between the snapshots. Of the deleted articles, we looked at what proportion of them exist in the 2008 snapshot to determine whether the deletion was persistent.

The results of this analysis are shown in figure 12, which shows the proportion of deletions that are persistent as a function of the article's age at the time it was deleted. We see a trend that shows deletions occurring early in an article's life are more likely to be persistent than deletions that occur later in an article's life. So, not only are articles unlikely to be deleted late in their life, but if a deletion does occur, it is less likely to be persistent. A common reason that a deletion is non-persistent is that the deletion was done for maintenance reasons that are tangential to whether the article is appropriate for Wikipedia. For example, an article might be deleted if a related article is being renamed to replace it.

These observations lead us to an answer to **RQ *Article Life Span***. Wikipedia's articles are here to stay, including those in its long tail. Once an article has survived the first few days of life, the chance that it is persistently deleted at some later date is small.

8. CONCLUSIONS

In each of the preceeding sections we gave a nuanced answer to one of the five research questions. Here we briefly summarize those questions and answers.

RQ *Long Tail Visits*: To what extent do Wikipedia viewers look at articles in the tail?

The visit distribution to articles in Wikipedia follows a log-normal curve. The top articles are by far the most popular, but the long tail accounts for a substantial fraction of visits to Wikipedia.

RQ *Wikipedia Growth*: How have article birth and mortality rates changed over time?

Wikipedia's article count continues to grow by thousands of articles per day. However, the birth rate is steady and the article mortality rate is slowly increasing, suggesting that the rate of growth has peaked and may begin declining.

[17]http://www.mediawiki.org/wiki/API

RQ *Topic Notability*: As time passes, are the articles that survive in Wikipedia increasingly on obscure topics?

Yes. New articles that are added to Wikipedia *are* increasingly on obscure topics as measured by our readership and search engine test metrics.

RQ *Deletion Reasons*: What are the reasons given for deleting articles? How do these reasons relate to the long tail?

The most common reason for deleting articles is "lack of notability". The use of the notability argument is evidence of resistance within the community to including articles that are arbitrarily far down the long tail of potential Wikipedia subjects.

RQ *Article Life Span*: When in the life of an article is it most likely to be deleted?

Most articles either have a very short life or a very long life. There is little evidence to date that the long tail is effectively being trimmed over time.

Analysis alone cannot resolve the debate about whether the diversity of "long tail" articles strengthens Wikipedia, or whether these obscure articles weaken its encyclopedic nature. This debate is over what determines the health of an online user-maintained encyclopedia. Since such encyclopedias have only existed for about seven years, it is no surprise that there is as of yet no clear answer.

Analysis can, however, help frame the debate. For instance, it is interesting that the probability of a new article being deleted has been increasing steadily over the past three years. The articles deleted for "lack of notability" that were analyzed using Yahoo! Search for this paper had average estimated notability less than that of surviving articles. However, since the estimated notability of newly created articles that survive has been declining in recent years, Wikipedia seems to have reached an intriguing inflection point: the articles that survive may be of comparable notability to those that are deleted. How will the conflict be resolved?

9. FUTURE WORK

The Wikimedia Foundation could enhance prospects for further research by making records of deleted articles available whenever appropriate. Combining this data with the viewership log data that is now becoming available will enable rich new analyses. Other research that would be interesting include an analysis of archived policy debates and inclusion/deletion discussions, a user-centric analysis of who is involved in these processes, and finally, quality and accuracy assessments of long tail articles. The latter is particularly important, because there is reason to predict that articles that are seldom viewed will have low quality on average.

10. ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the members of GroupLens Research, especially the Wikipedia Crew. We thank the Wikimedia Foundation for their support of research by sharing a 1/10th feed of their web logs with us. This work is funded by the National Science Foundation, grants IIS 03-24851, 05-34420, 07-29286, and 08-08711.

11. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *Proc. WSDM 2008*, pages 207–218, Palo Alto, CA, USA, 2008. ACM.

[2] C. Anderson. *The Long Tail: Why the Future of Business is Selling Less of More*. Hyperion, July 2006.

[3] I. Beschastnikh, T. Kriplean, and D. W. McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *Proc. ICWSM 2008*, Seattle, WA, USA, 2008. AAAI.

[4] S. L. Bryant, A. Forte, and A. Bruckman. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *Proc. GROUP 2005*, pages 1–10, Sanibel Island, FL, USA, 2005. ACM.

[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *0706.1062*, June 2007.

[6] A. Forte and A. Bruckman. Scaling consensus: Increasing decentralization in Wikipedia governance. In *Proc. HICSS 2008*, page 157, 2008.

[7] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, Dec. 2005.

[8] S. A. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. *J. Inf. Sci.*, 32(2):198–208, 2006.

[9] M. L. Goldstein, S. A. Morris, and G. G. Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B - Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

[10] A. Kittur, E. H. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Proc. CHI 2007*, Montreal, Quebec, Canada, 2007. ACM.

[11] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *Proc. CHI 2007*, pages 453–462, San Jose, CA, USA, 2007. ACM.

[12] T. Kriplean, I. Beschastnikh, D. W. McDonald, and S. A. Golder. Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *Proc. GROUP 2007*, pages 167–176, Sanibel Island, FL, USA, 2007. ACM.

[13] D. Milne, O. Medelyan, and I. H. Witten. Mining domain-specific thesauri from Wikipedia: A case study. In *Proc. WI 2006*, pages 442–448. IEEE CS, 2006.

[14] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in wikipedia. In *Proc. GROUP 2007*, pages 259–268, Sanibel Island, FL, USA, 2007. ACM.

[15] THEwikiStics. Yearly wikimedia page hits comparison. http://wikistics.falsikon.de/2008/, 2008.

[16] F. B. Viegas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. CHI 2004*, pages 575–582, Vienna, Austria, 2004. ACM.

[17] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *Proc. WSDM 2008*, pages 171–182, Palo Alto, CA, USA, 2008. ACM.

[18] D. M. Wilkinson and B. A. Huberman. Cooperation and quality in Wikipedia. In *Proc. WikiSym 2007*, pages 157–164, Montreal, Quebec, Canada, 2007. ACM.

[19] F. Wu, R. Hoffmann, and D. S. Weld. Information extraction from Wikipedia: Moving down the long tail. In *Proc. KDD 2008*, pages 731–739, Las Vegas, NV, USA, 2008. ACM.