Wikipedia: Nowhere to grow

Austin Gibbons, David Vetrano, Susan Biancani

11 June 2012

Wikipedia is a free, collaboratively edited, multi-lingual encyclopedia, founded in 2001. Today, it has grown into a massive effort to collect and categorize human knowledge in all of the world's active languages. In its first several years of existence, English Wikipedia grew rapidly, both in number of articles and in number of editors. Researchers characterized the growth rate of wikipedias as exponential and identified a self-similar mechanism of growth (Almeida et al. 2007, Spinellis and Panagiotis 2008, Ingawale and Dutta 2009). However, since 2007, the growth of English Wikipedia has slowed, with fewer new editors joining, and fewer new articles created (stats.wikimedia.org).

While several mechanisms have been proposed to explain this slow-down, we believe an important one remains largely unexplored: that the larger the site becomes, and the more knowledge it contains, the more difficult it becomes for editors to make novel, lasting contributions. That is, all of the easy articles have already been created, leaving only more difficult topics to write about. We call this the Low-Hanging Fruit hypothesis. This paper is organized as follows: we will explain the background and related work, our hypotheses, our data and data-management methods, several experiments addressing our hypotheses, discussion of our findings and suggestions for future work, and finally, recommendations for the wikipedia community.

1 Background and Related Work

Although the slow-down in Wikipedia's growth has been multi-faceted, the decrease in the number of new editors joining has perhaps received the most attention. Wikipedia's parent, the Wikimedia Foundation, has sought to investigate this question, examining the retention rate of new editors (those with less than one year experience) versus more experienced editors (those with more than one year experience) (`http://strategy.wikimedia.org/wiki/Editor_Trends_Study/Results`). They find a precipitous decline in the retention of new editors from mid-2005 to mid-2007, which accounts for much of the change in the number of active editors. Retention of experienced editors decreased, but much less dramatically, during this time. Since 2007, editor retention rates have appeared to be more stable.

Several mechanisms have been proposed to explain the observed slow-down. These include: an unfriendly or closed atmosphere, with newer users' edits especially likely to be reverted (Halfaker et al. 2011); the related possibility that new editors are more likely to be incompetent or acting in bad faith (Halfaker 2012); negative perceptions of the type of person who becomes an editor, specifically, that editors are "geeky," "nerdy," and "unkempt, unhealthily obsessive, and absorbed with online life" (Antin 2011, p. 3416); increased overhead costs to coordination and production, and the possibility that Wikipedia has reached the natural limit of its growth (Suh et al. 2009). While we agree that all these factors may be operating, in this paper, we elaborate on the last of them.

Suh et al. hypothesize that Wikipedia editors face increasingly limited opportunities to make novel contributions to the site, giving rise to increased conflict. They argue that two types of factors determine Wikipedia's capacity for growth: internal limits, such as the number of available volunteers, the hours the volunteers can spend, and their motivation for the work; and external factors, such as the amount of publicly available and relevant knowledge that editors can access and report, and the usability and functionality of the tools editors and administrators use to do their work. The authors address their questions by dividing the population of Wikipedia editors into four classes, based on their activity level, and then examining changing patterns in the types of ac-

tivities each class engages in over time. They do not directly investigate their hypothesis that Wikipedia is limited by the amount of publicly available knowledge, or seek to determine the effect that this limit has on Wikipedia's growth. Here, we aim to do just that.

As Suh et al. argue, "In earlier days, a group of non-specialist volunteers, armed with a search engine, were able to create and edit pages with little time and effort" (p. 9). Today, those articles are already made, and new editors must seek out increasingly specialized topics in which to contribute. In this respect, Wikipedia has followed a similar pattern to academia and technology: Jones (2009) finds that scientific fields with deeper knowledge bases show greater levels of specialization, and as a consequence, higher rates of teamwork. As it becomes necessary for an innovator to have an ever more sophisticated technical background, it becomes difficult to produce novel work on one's own; Jones argues that we see "the death of the renaissance man." We argue that a similar dynamic is at work in Wikipedia: in order to make a novel, useful contribution to the site, editors must meet an increasingly high bar of expertise in their field. As this bar rises, the pool of available, qualified editors shrinks. We hold that this shrinking pool explains much of the slowdown in Wikipedia's growth.

2 Hypotheses

Our argument leads us to the following hypotheses:

1. The slowdown in growth should be observed across all or most of the different language-based wikipedias. Because these sites were created at different times, and have a different numbers of editors, we do not expect that all will slow down concurrently; however, we do expect that all will show similar, plateau-shaped patterns of knowledge saturation.

2. Older articles are more accessible. They will be more popular to edit than more newly created articles.

3. Older articles (those created earlier) have broader appeal. They will be more popular to read than more newly created articles.

We test these hypotheses in a series of experiments on Wikipedia data from several languages.

3 Data and Data-Management Methods

Wikipedia provides nearly all of its data publicly on dumps.wikimedia.org, with some user details anonymized. Most of the data is available as compressed XML. The data on all languages other than English were fetched in early May 2012. From these data dumps we pulled the number of views each page received in January and February 2012, the complete revision history of every language (excluding English), and the administrative logs (such as adding and deleting pages) for every language. English, being the largest wikipedia by around an order of magnitude, had a pre-processed dataset called DiffDB which was a by-product of the Wikipedia Summer of Research and allowed us to collect information about the differences between successive revisions.

All large-scale computation was done using Amazon's Elastic MapReduce (EMR), using streaming Hadoop. Specifically, EMR is a platform which abstracts away many of the details of setting up and configuring the instances, requiring only rare modifications to the job configuration. Datasets were placed in compressed form in S3 buckets, from which Elastic MapReduce then directly reads and to which it writes. Decompression of the compressed files is done transparently. A mapper and reducer are then written; from the programmer's perspective, the mapper and reducer read tab-separated lines of text from standard in and output tab-separated lines of text to standard out. All input data will be sent to exactly one mapper and all lines with the same key will be sent to the same reducer.

The paradigm used by streaming Hadoop can be quite powerful. Allow us to consider an example set of map-reduce steps which was used to bin pageviews into buckets by the time of first touch by a user. First, stubs, which contain meta information about every revision were processed for each language. A mapper streamed through the XML, collecting data about all of the revisions for an article. Using our bot detection strategy, the title of the article in URL-encoded form was recorded along with the first edit by a non-bot user. In this step, the identity reducer was used. Next, pageviews were processed in the same manner, outputting lines of article name. Finally, a third map-reduce task was launched to join these two datasets on article name, using the trick of labeling pair types and sending each to the same reducer with the article name as a key. We were then able to analyze the

aggregated page view statistics on our local machine.

A large number of revisions are made by bots. This pattern is especially true in smaller languages, where bots can comprise in excess of 80% of the total number of revisions. Wikipedia provides several lists of bots; however participation in these lists is optional, and we have found it not to be comprehensive. To mitigate this issue, we used a simple strategy: ignore all revisions whose user name contains the string "bot." While this may ignore some users spuriously, we feel that these users will be a roughly unbiased sample of the population, will not tend to have any special properties, and so can be ignored. In all analyses below, we have filtered out activity by bots, unless otherwise noted. In addition, we have included only data from Wikipedia Namespace 0, which is the namespace used for articles (as opposed to talk pages and the like).

4 Preliminary Work

4.1 Clustering Editors

Our initial hypothesis, following existing work (Halfaker et al. 2011), was based on the idea that there was a hostile climate that was both driving away existing editors and discouraging potential new users. To test this hypothesis, we attempted to cluster users into different roles and observe how those roles changed over time, both in size and in the individuals belonging to those clusters. We hoped to identify a cluster of malicious, hostile, or otherwise detrimental users, and observe it growing in conjunction with editor retention rates. We built feature vectors for registered users which included information about the number of additions and deletions they contributed to a wikipedia, amount of text added and deleted, the total size of comments, and the number of pages they edited. We used k-means clustering to group the editors, and did this for every month, using the previous six months as the history for that month. For several attributes, the long-tail necessitated moving the data into log-space prior to clustering. After experimentation, we found that the clustering was largely unstable, and that different initialization points would lead to very different clusters. We were able to identify some factions of users, but these were not very consistent and did not offer much insight about the problem we were addressing.

4.2 Inteviews

As a result, we decided to reach out to the Wikipedia community to understand how the editors themselves perceived their community. In total, we spoke with five editors who had each made over ten thousand edits: two through personal contacts, two through social media, and one via cold contact. One of the editors described Wikipedia as "the most difficult community on the Internet," and helped motivate our initial foray into analyzing the hostility. The latter four offered weight to our second hypothesis; they generally acknowledged that there were occasional arguments on the internal talk pages, but did not believe that this explanation was the the primary factor in declining editor rates. One editor suggested that the effect was because the newer articles being created and edited were less interesting to the general population and required more specialized knowledge. Following this thread, and after initial exploration of the data, we formed a new hypothesis, deemed the "low-hanging fruit" theory. We hypothesize that broad-interest, accessible articles - those which require little or no domain expertise - are created earlier in a wikipedia's lifetime, edited more, and viewed more than other articles. When these accessible articles have been created and revised, many users may lose interest in further editing the wikipedia.

5 Experiments

5.1 Data selection

To demonstrate the low-hanging fruit hypothesis, we analyzed editor trends across many different languages. We chose languages that represented small, affluent, well-educated, geographically centralized populations such as Japanese, Korean, Finnish, Hungarian, Norwegian, and Estonian. In addition, we selected a few larger languages, with more geographically dispersed user bases, including Spanish, Russian, and Portuguese. We further believe that the number of individuals who edit articles in more than one of these wikipedias is trivially small and so can be safely ignored.

Figure 1 illustrates basic descriptive statistics about these languages. The horizontal axis shows when the language was started. Specifically, it indicates the date at which the language first reached 5% of its current size, in count of articles. The vertical axis shows current size, in count of articles. Finally,

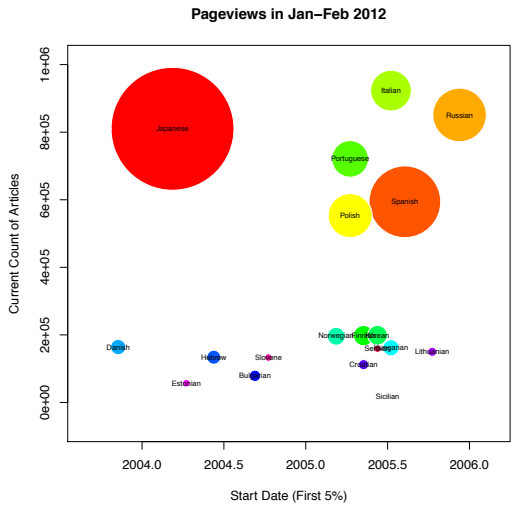

Figure 1: Descriptive statistics on included Wikipedias

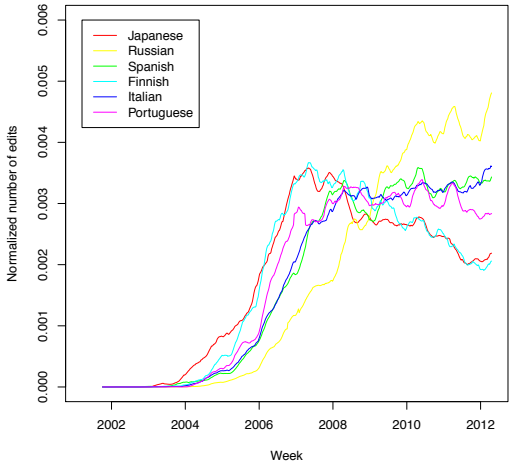

Figure 2: Total edits per week normalized by size of wikipedia

the area of each circle corresponds to the total count of pageviews of all pages in the Wikipedia site for that language, received in January and February, 2012.

5.2 Test of Hypothesis I

The slowdown in growth should be observed across all or most of the language communities that create Wikipedia.

We first demonstrated that all of these wikipedias observed a slowdown in their growth. We computed the number of edits that occurred in each week of that wikipedia's existence, and smoothed over a sixteen week interval. Figure 2 displays the weekly count of edits in each of six languages: the x-axis shows the week in which the edits were made, and the y-axis shows the count of edits to all pages in that language in that week. We quickly observe that all of these languages experience a period of exponential growth followed by a plateau or decline in activity. While the order in which languages reached a plateau bears some resemblance to the order in which they were founded (with Japanese among the first and Russian last), it does not line up perfectly. Even so, this plot confirms that plateaus in growth can be observed in many wikipedias, not just English.

To further investigate these plateaus, we aligned our data on a different time-axis. We looked at the rate of edits in each week after the language reaches 5% of its current size in articles. This is seen in Figures 3, 4, 5. The data is smoothed on an 80 week moving average for legibility. The y-axis shows the count of edits per week, normalized by the total number of edits in that language.

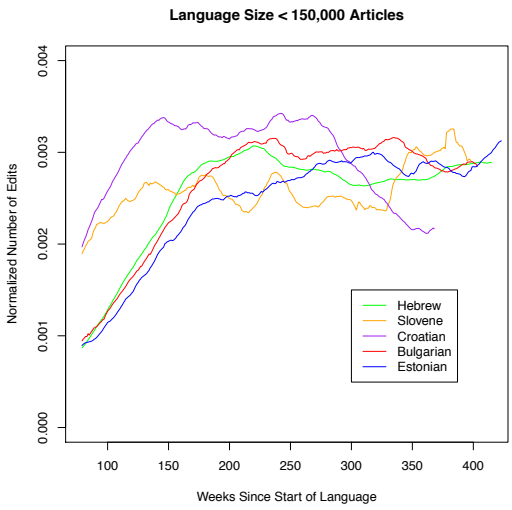

Figure 3: Normalized Count of Edits by Age of Language - Small Languages

These figures, which are divided according to the current size of each language, show that all of these languages experience a period of rapid growth followed by a plateau period. Many of these languages have behavior which is very closely aligned, such as Portuguese, Italian, and Spanish, whereas a few (Croatian, Korean and Russian) do not show evidence of a plateau. However, the prevailing pattern is that of a slowdown in growth. Because this plateau

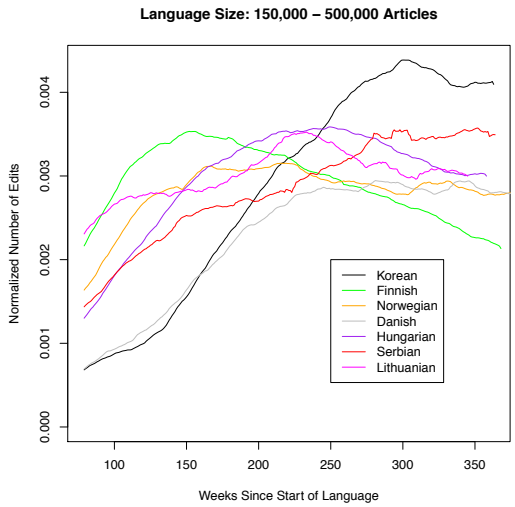

Figure 4: Normalized Count of Edits by Age of Language - Medium Languages

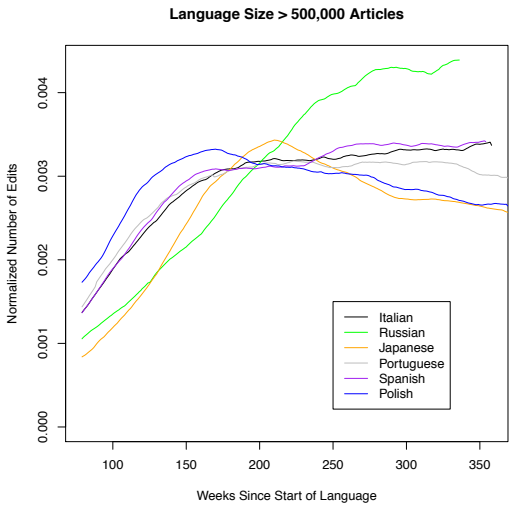

Figure 5: Normalized Count of Edits by Age of Language - Large Languages

pattern is not specific to one site or to one community of editors, we think it unlikely that it is driven by the hostility of a single group of editors. Moreover, because these plateaus occurred at different dates in real time, we also think it unlikely that any secular effect - such as the creation of some other popular website, or a change in policy at Wikipedia that affected all sites - is responsible. We think it will be interesting in the future to investigate the variance in growth rates: do Russian, Korean, and Croatian share important characteristics that set them apart from the other languages?

5.3 Test of Hypothesis II

Older articles (those created earlier) will be more popular to read than more newly created articles.

To demonstrate our claim that articles more recently created are less interesting to edit, we look only at the edit history of the year 2011 with regard to what year the article being edited was created. We normalize by the total number of edits in the year 2011 for each language. The results are shown in Figures 6, 7, and 8. We observe a striking effect:

All of these languages follow the same trend: editors are primarily interested in editing articles created in this year. Editors are next interested in editing articles that were created during the infancy period of a wikipedia, preferring to edit these articles by a factor of two over articles created in the intermediate period. Once again we can notice that the more mature a language is, the more pronounced is this effect, with Japan peaking earlier and slumping lower than Russian. This demonstrates that editors show a preference for editing articles created earlier in a wikipedia's development.

Figure 6 contains the largest languages we observed. When we look at the wikipedias from smaller language communities, we observe that this trend does not hold. This may fit with our hypothesis, if these languages are still picking the low-hanging fruit. These plots suggest that overall size plays a mediating role in the time-dynamics we have observed. Although most of these small languages showed a plateau in overall editing activity similar to that of the larger languages (see above), their editors do not seem to preferentially edit older articles.

We further investigated this by asking whether the creation date of articles is associated the number of editors who work on them. Using data only from English Wikipedia, for each week, we collected all the articles created in that week, and then tabulated the number of unique editors who touched each article in the first year after its creation. We have plotted the mean number of unique editors per article by week of article creation in Figure 9; the data have been smoothed in a six-week moving window. We are not sure what to make of the spikiness of seen in 2001-2003; perhaps the data is more volatile here because there are many fewer articles than later on. From 2006-2011, there is a fairly steady decline in the mean number of edi-

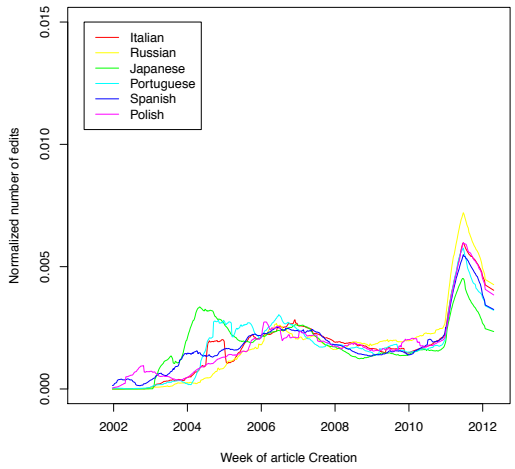

Figure 6: Edits in 2011 normalized by size of wikipedia

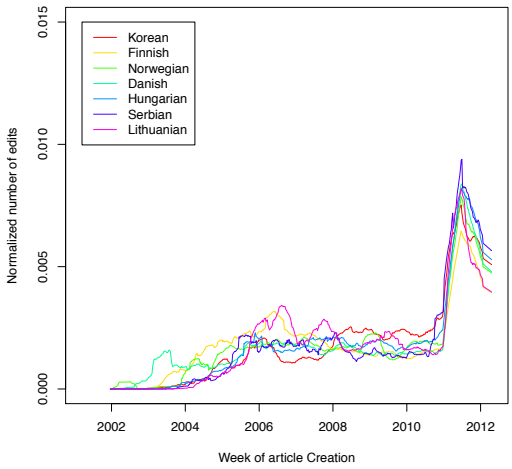

Figure 7: Edits in 2011 normalized by size of wikipedia

tors: articles created in 2006 have an average of 4.5 editors in their first year, while those created in 2010 have an average of about 2.6. This trend occurs despite the fact that Wikipedia has grown, in absolute numbers of active editors, from about 100,000 active editors in June 2006, to about 700,000 in June 2011 (http://stats.wikimedia.org/EN/ TablesWikipediaEN.htm). Until we further investigate the volatility in editor popularity in 2001-2003, we are not confident basing strong conclusions on this graph, but in the meantime, it provides suggestive evidence worth additional study.

5.4 Test of hypothesis III

Older articles will be more popular to read than more newly created articles.

In addition to the patterns we observed in revisions across the languages, we investigated pageview data as well. Because we argue that articles with broad appeal were created first, we expect these older articles to receive more pageviews. We constructed a graph of pageviews binned into buckets by date of first touch by a human. Figure 10 shows the count of total pageviews in January and February, 2012 for all articles created in a given week. The majority of pageviews are to articles from the early days of Wikipedia. In fact, aggregating data across all languages on Wikipedia, only 2.6% of the page views in January/February 2012 were for pages created in 2011, while pages created in 2002, 2003, and 2004 respectively garnered 17.9%, 14.1%, and 19.3% of the page views during this time period. These findings echo those by Lam and Reidl (2009), using pageviews of English Wikipedia from October 2007-December 2007. The limitation on both our data and theirs is that is difficult to control for aging effects: the longer an article exists, the more time it has to accumulate in-links, both from other articles in Wikipedia (i.e. backlinks) and from the web at large. These effects may account for some of the differences between older articles and newer ones.

Lam and Reidl attempt to control for this effect by grouping articles by the number of backlinks they have, and comparing within groups. Doing so, they find that the effect still holds, though to a lesser extent. The authors conclude that links to articles play a role in promoting pageviews but do not completely explain the differences in popularity between articles created early and those created late. We follow Lam and Reidl in arguing that the data as they stand suggest that pages created early in the history of wikipedia are different from those created more recently, appealing to a broader audience. This challenges the often-tacit assumption that the nature of the remaining work on Wikipedia is the same as during the early days.

6 Discussion

We have investigated three related hypotheses regarding the slowdown in Wikipedia's growth: (1) a

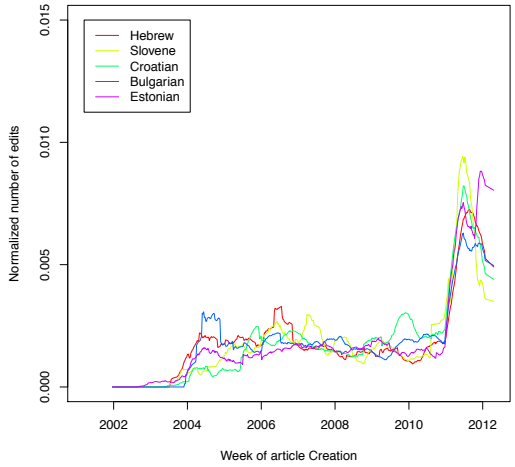

Figure 8: Edits in 2011 normalized by size of wikipedia

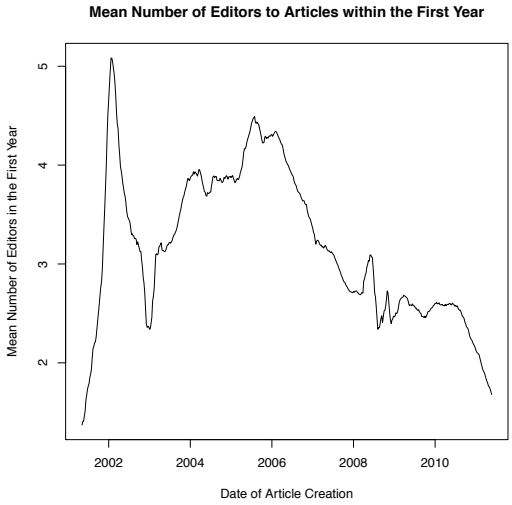

Figure 9: Editors per article in in the first year, by creation date of article

similar slowdown is observed across many languages with diverse characteristics; (2) articles created earlier are more popular to edit; and (3) articles created earlier are more popular to view. We have found support for all three hypotheses, lending weight to our contention that the exhaustion of the "low-hanging fruit" of Wikipedia editing (i.e. the creation and editing of accessible, broadly appealing articles) plays a key role in explaining Wikipedia's slowing growth.

We see several avenues for further work on this question. West et al. (2012) use data from the Yahoo! Toolbar to show that editors of Wikipedia have expertise in the areas they edit. Unfortunately, their data begins in 2008, and so cannot be used to study the earliest period of Wikipedia's growth. However, it may be informative to examine patterns from 2008-2011, to ask whether we see evidence of increasing expertise or specialization of editors over this time period. Our preliminary data in Figure 9 show a steady decline from 2006-2011; we are interested to know if the data from Yahoo! Toolbar corroborates this decline over the years it covers.

Iba et al. (2010) study, among other things, the variation in the networks of editors of different wikipedia articles. They show that, for an accessible article, such as "Australia", the graph of editors (in which there is a tie from A to B if A's edit is immediately followed by B's edit) contains many nodes, has long average path lengths, and a large diameter. In contrast, the edit graph for a specialized article, such as "Mozart in Italy" has fewer nodes, short average path length, and a small diameter. We have attempted to investigate similar questions by looking at the number of unique editors on each article, but it may help to use a more sophisticated analysis, similar to that used by Iba et al.

We also think it would be informative to borrow the approach Lam and Reidl (2009) used to control for the number of backlinks to each article, in their investigation of the notability of older versus more recent articles. These authors point out that among Wikipedia editors, there are two camps: the inclusionists, who believe the encyclopedia should be as comprehensive as possible, and the deletionists, who believe it should focus on noteworthy articles to the exclusion of more obscure topics. If the inclusionist viewpoint dominates, we will see more articles created that have few pageviews and few editors; this may account for the trends we have described above. We could investigate this question by setting some threshold, N, and considering only articles that have more than N pageviews per month. Of the articles that have more than N pageviews per month today, what is the distribution of their creation dates? Similarly, of the articles that have more than M editors per month, what is the distribution of their creation dates? This approach would allow us to filter out the very obscure articles that would be created under a highly inclusionist regime, and to isolate the time trends in the creation of popular articles.

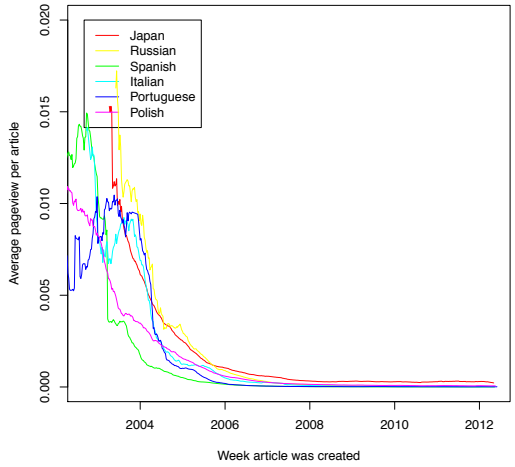

Figure 10: Mean pageviews per article, by creation date of article

7 Recommendations

Our recommendation to the Wikimedia Foundation is to first confirm these findings independently, and reach out to its editing community for feedback. We further recommend the Wikimedia Foundation embraces the cultural shift that a mature wikipedia requires, and devote more resources to recruiting editors with specialized knowledge. In particular, we recommended expanding the Wikipedia Campus Ambassadors program to a more aggressive position, and to encourage students to translate their class work into contributions to Wikipedia, as well as working with University professors and high school teachers in promoting Wikipedia among their advanced students.

8 Conclusion

We have presented a range of evidence supporting the hypothesis that much of the plateau in Wikipedia's growth can be attributed to the exhaustion of topics that are broadly accessible and popular to read and edit. The evidence presented spans many languages and many independent communities of editors. While it is not conclusive, we believe it provides compelling, early-stage support for our thesis, and we look forward to doing continued work in this area.

9 References

- Almeida, Rodrigo B, Barzan Mozafari, and Jungchoo Cho. 2007. "On the Evolution of Wikipedia." in ICWSM. Boulder, CO.
- Antin, Judd. 2011. "My Kind of People ? Perceptions About Wikipedia Contributors and Their Motivations." Pp. 3411-3420 in CHI Session: Inventives and User Generated Content. Vancouver, BC, Canada: ACM.
- Halfaker, Aaron, Aniket Kittur, and John Riedl. 2011. "Don í t Bite the Newbies : How Reverts Affect the Quantity and Quality of Wikipedia Work." Pp. 163-172 in WikiSym. Mountain View, CA: ACM.
- Iba, Takashi, Keiichi Nemoto, Bernd Peters, and Peter a. Gloor. 2010. "Analyzing the Creative Editing Behavior of Wikipedia Editors." Procedia - Social and Behavioral Sciences 2(4):6441-6456. Retrieved March 6, 2012 (http://linkinghub.elsevier.com/ retrieve/pii/S1877042810011122).
- Ingawale, Myshkin, Rahul Roy, Amitava Dutta, and Priya Seetharaman. 2009. "The Small Worlds of Wikipedia: Implications for Growth , Quality and Sustainability of Collaborative Knowledge Networks The Small Worlds of Wikipedia : Implications for Growth , Quality and Sustainability of Collaborative Knowledge Networks." in Proceedings of the Fifteenth Americas Conference on Information Systems. San Francisco, CA: AMCIS.
- Lam, Shyong (Tony) K., and John Riedl. 2009. "Is Wikipedia growing a longer tail?" P. 105 in Proceedings of the ACM 2009 international conference on Supporting group work - GROUP í09. Sanibel Island, FL: ACM Press Retrieved (http://portal.acm.org/citation.cfm?doid=1531674.1531690).
- Ransbotham, Sam, and Gerald C Kane. 2011. "Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia." MIS Quarterly 35(3):613-627.
- Suh, Bongwon, Gregorio Convertino, Ed H Chi, and Peter Pirolli. 2009. "The Singularity is Not

Near : Slowing Growth of Wikipedia." in WikiSym, vol. 1. Orlando, FL: ACM.

- West, Robert, Ingmar Weber, and Carlos Castillo. 2011. "Smart but Fun : A Data-Driven Portrait of Wikipedia Editors." (November).