The Singularity is Not Near: Slowing Growth of Wikipedia

Bongwon Suh, Gregorio Convertino, Ed H. Chi, Peter Pirolli
Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA, 94304
+1 (650)812-4806
{suh, convertino, echi, pirolli}@parc.com

ABSTRACT

Prior research on Wikipedia has characterized the growth in content and editors as being fundamentally exponential in nature, extrapolating current trends into the future. We show that recent editing activity suggests that Wikipedia growth has slowed, and perhaps plateaued, indicating that it may have come against its limits to growth. We measure growth, population shifts, and patterns of editor and administrator activities, contrasting these against past results where possible. Both the rate of page growth and editor growth has declined. As growth has declined, there are indicators of increased coordination and overhead costs, exclusion of newcomers, and resistance to new edits. We discuss some possible explanations for these new developments in Wikipedia including decreased opportunities for sharing existing knowledge and increased bureaucratic stress on the socio-technical system itself.

Categories and Subject Descriptors

H.5.3 [Information interfaces and presentation]: Group and Organization Interfaces—Web-based interaction, Collaborative computing, Evaluation/methodology; K.4.3 [Computers and society]: Organizational Impacts—Computer-supported collaborative work

Keywords

Growth, power law, logistic model, Wikipedia, resistance, population

1. Introduction

Many natural systems have a fundamental growth process that leads one to expect exponential (or geometric) growth rates, so long as the process does not eventually become limited by some other constraint. Biological populations, when unconstrained, tend to exhibit such growth. This exponential growth happens, for instance, with microbes that repeatedly split into two daughter microbes. Many measures of digital systems have historically tended to exhibit such exponential growth, including the number of transistors in an integrated circuit (i.e., Moore's Law [11]), memory capacity, and the amount of content stored in digital media [19]. Given existing growth trends for digital systems, it is not surprising that several papers e.g., [24] and [29] have suggested that Wikipedia shows such exponential growth and that growth is mainly spurred by exponential growth in contributing editors [2].

The existing trends of exponential growth in digital technologies were the basis for Kurzweil's [17] argument that biological evolution and technological evolution follow a law of accelerating returns (i.e., exponential or even super-exponential growth). This lead to the notion of the "Singularity": a point in the near future when technological change becomes "so rapid and profound that it represents a rupture in the fabric of human history."[1] We argue that Wikipedia, one of the world's largest knowledge aggregators, does indeed mirror the growth of natural populations, but, following Darwin [7], we suggest that this growth becomes increasingly constrained and limited, and under those conditions there will be increased evidence of competition and dominance.

In this paper, we present data that challenges the notion that Wikipedia exhibits unconstrained exponential growth in editor participation and contribution. We will show that growth has decreased substantially over the last two years, perhaps indicating some fundamental limiting constraints to growth. In ecological systems, when unfettered population growth approaches natural limits (e.g., in available resources), one generally observes increased competition. For Wikipedia, we will examine the data for indicators of increased competition that would be expected as a growing population system comes up against limits to growth. We present data from Wikipedia addressing three different aspects over time: the global activity level, a detailed analysis of the edit rates of various editor classes, and the population shifts in editor classes.

2. Background and Related Work

Started in 2001, Wikipedia is now the largest public collection of encyclopedic knowledge in the world. It is a complex socio-technical system formed by volunteers who collaboratively edit content via a wiki. It is regulated by policies and roles that are defined collectively by the community. Wikipedia has been a subject of considerable interest from researchers because of its rapid rise, its apparent capacity to aggregate all encyclopedic knowledge, and as an exemplar of a novel form of distributed peer-production system that yields large and finely detailed data sets concerning its evolution and dynamics. Wikipedia has become a kind of "living laboratory" for the study of the production and sharing of content by online communities [6].

One thread of such research has applied stochastic models [30] and network-theoretic models (e.g., [24] and [37]) to explain strong regularities in Wikipedia content production. For instance, the distribution of number of edits per page as a function of rank

WikiSym'09, October 25–27, 2009, Orlando, Florida, U.S.A.

[1] http://www.kurzweilai.net/articles/art0134.html

order is predicted by a simple stochastic growth process combining an "edits beget edits" process plus a decay process in editing rate due to page age [30]. A preferential attachment model has been proposed to account for the way that new articles "sprout" from, and reference, old articles [24]. Such a model does not preclude some ultimate limitation to growth, although at the time it was presented [24] there was an apparent trend of unconstrained article growth. Other models and empirical research have claimed that Wikipedia article growth is exponential because there is an exponential growth in the number of editors contributing to Wikipedia [2].

On the other hand, statistics reported in Wikipedia itself[2] suggests that the rate of growth of Wikipedia pages has declined since 2007. This is suggestive of a growth process that is coming against resource limits. In natural populations, as population growth comes up against limits of an ecological niche (e.g., in available space or available energy), competition increases, and advantages go to the members of the population who have competitive dominance. By analogy we hypothesize that:

(a) that the population of editors contributing to Wikipedia is increasingly facing limited opportunities to make novel contributions, and

(b) the consequences of these (increasing) limitations in opportunities will manifest itself in increased patterns of conflict and dominance.

Our overall goal is to develop an updated growth model of Wikipedia over time. To do this, we need to understand the various types of editor behaviors, and how they affect overall growth patterns.

Some researchers found that there is a drastic inequality in the editors' contribution to Wikipedia. Priedhorsky et al. [22] measured the relationship between editors' edit count and the editors' ability to convey their writings to Wikipedia readers, measured in terms of persistent word views -- the number of times a word introduced by an edit is viewed. The researchers analyzed 25 trillion persistent word views attributable to registered users between 2002 and 2006. The study result shows that the top 10% of editors (by edit count) were credited with 86% of persistent word views (PWV), the top 1% about 70%, and the top 0.1% (4200 users) were attributed 44% of PWVs, i.e. nearly half of Wikipedia's "value" as measured in this study.

In our own research, Kittur et al. [15] analyzed the entire edit history of Wikipedia up to July 2006 and reported that the influence of administrators on content production has steadily diminished since 2003. The paper reports that administrators performed roughly 10% of the edits in 2006 while they contributed 50% of total edits in 2003. This happened despite the fact the average number of edits per administrator had increased more than fivefold during the same period. While these two papers provide an interesting perspective on how contents are generated, the results of the studies provide an outdated view on Wikipedia as of mid-2006 before the slowdown of Wikipedia began to occur, as we will show below.

Here we aim to provide an updated analysis of Wikipedia growth patterns, including how different types of editors contribute at different rates, and are experiencing differing levels of resistance.

[2] http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth

3. Activities, Actors, and Processes

In this section, we provide an overview of Wikipedia as a collaborative authoring environment aimed for the entire global community to co-create an encyclopedia for the Web.

Wikipedia is a free, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Wikipedia's 12 million articles (2.8 million in the English Wikipedia) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone who can access the Wikipedia website [35].

3.1 Activities in Wikipedia

The primary activity in Wikipedia is editing of the content. Users can create a new page, add/modify/remove contents in existing pages. It should be noted that Wikipedia contains many different types of pages other than the content articles or simply articles. For example, the "**discussion**" pages are associated with each article and are used to coordinate work among multiple editors.

Figure 1: Each Wikipedia article has links to its discussion page and edit history. For registered users, Wikipedia shows addition links such as "my watchlist", "my talk", etc.

Other than content pages and associated discussion pages, there are pages designated to facilitate communication between users. Registered users have a **user page** in Wikipedia (for example, "User:JohnDoe") that can be used to present oneself, for project-related bookmarks, and for drafts, tests, and other working material. User pages also have associated **talk** pages (for example, "User talk:JohnDoe"). Those pages are also intended for discussion between Wikipedia users. When one editor needs to contact another editor, it is typical that one leaves a message on the receiver's talk page. Any registered user will be notified when someone leaves a message that way, with a notice the next time the user logs in to Wikipedia.

In addition, as shown in Figure 1, the "**history**" page, which is also attached to each article, records every single past revision of the article. This feature enables Wikipedians to easily compare old and new versions, undo changes that an editor considers undesirable, or restore lost content. Regular contributors often maintain a "**watchlist**" of articles of interest to them, so that they can easily keep tabs on all recent changes to those articles.

Vandalism is any addition, removal, or change of content made in a deliberate attempt to compromise the integrity of Wikipedia. When spotted, an editor can repair vandalism by reverting the changes. Wikipedians also warn the user who committed the vandalism. Users who vandalize Wikipedia repeatedly in spite of the warning are reported to administrators and blocked from making any further edits. In an early study, Viegas et al. [27] found that vandalism is repaired rather quickly in Wikipedia.

3.2 Actors in Wikipedia

Editors of Wikipedia can register themselves and create an account while non-registered users also can edit pages. Editors in good standing in the community can run for one of many of levels of volunteer stewardship. The most notable of all is "**administrator**", a group of privileged users (1,625 Wikipedians for the English edition on February 21, 2009), who have the

ability to perform administrative roles such as delete pages and block users from editing. Despite the name, administrators are not supposed to have any special privilege in decision-making and are prohibited from using their powers to settle content disputes [31]. The role of administrators is often described as "janitorial".

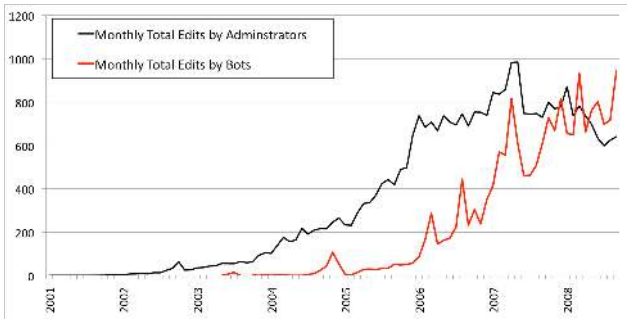

Figure 2: Monthly edits by administrators and *bots* (in thousands)

Computer programs called ***bots*** have been used widely to reduce humans' labor for managing Wikipedia pages. Bots are automated or semi-automated tools that carry out repetitive and mundane tasks such as removing vandalism, correcting common misspellings, and stylistic issues. There are currently 895 *bot* tasks approved for use on the English Wikipedia, however they are not all actively carrying out edits.

3.3 Processes, Guidelines, and Rules

As Wikipedia grows, the community has developed a swelling collection of **policies** (must be followed, no exceptions) and **guidelines** (should be followed, suggested) that regulate the editors' behavior. These norms are derived from a small list of general **principles**. Some define the standards for acceptable content: Neutral point of view, Verifiability, No original research, Biographies of living persons, Naming conventions (or style manual, recently added). Others define standards for acceptable behavior: Civility, No personal attacks, Harassment, No legal threats, Consensus, Dispute resolution. Policies or guidelines have been evolving [9] and typically discussed on the associated talk pages.

The founding policy principle has always been keeping the community **as rule-free as possible** (Ignore all the rules policy). However, Butler et al. [4] found that the amount of bureaucracy has rapidly grown in Wikipedia, together with the efforts involved (e.g., the talk page linked to the Ignore all the rules policy itself swelled over 3600%). All policies studied grew enormously in terms of word counts. This growth shows that policy development, discussion, and maintenance have become also an important part of the work of administrators or community facilitators.

Administrators in Wikipedia usually enforce these policies. They have additional access to restricted technical features to help with maintenance. They can **protect** and **delete** pages. When a page is protected, editing on the page is prohibited (partially if semi-protected). Also, administrators can **block** users from making contributions. Blocks sometimes are used as a deterrent, to discourage whatever behavior led to the block (e.g. vandalism, personal attack) and encourage productive editing. Administrators are also capable of undoing such actions; unprotect pages, restoring deleted pages, and unblocking users.

Consistent and reasonable enforcement of these policies and guidelines is difficult, and can result in novice users experiencing differing levels of resistance. A press article in 2008 [18] started identifying the risk that that novices in Wikipedia could quickly get lost in *bureaucracy*. The article cites estimates from 2006 suggesting that the entries about governance and editorial policies are one of the fastest-growing areas of the site and represent around one-quarter of its content.

3.4 Issues and Challenges

Despite its phenomenal success, Wikipedia also has been criticized for the un-authoritative and unreliable sources of information due to being open for editing by everyone [8]. Criticism levied includes inaccurate information, a non-neutral point of view and conflicts of interest [15], and vandalism [27][34]. The Wikipedia community also has been criticized for the systemic bias in coverage of topics, lack of credential verification [33], anti-elitism as deterrent for experts, and the "hive mind" consensus building [14].

Wikipedians have been aware of these issues and the Wikipedia community has been trying to develop procedures to improve the reliability of Wikipedia. The English-language Wikipedia has introduced an assessment scale against which the quality of articles is judged [32]. Since May 2008, the German-language Wikipedia implemented "flagged revision" system where a stable version of an article is shown until established Wikipedia editors confirm the latest edit as a clean version.

On the other hand, some researchers addressed the issue by designing tools to enhance the transparency in Wikipedia [1][26]. Providing tools and infrastructures mechanisms that support this type of work is an important requirement for building successful large-scale communities. Adler et al. [1] designed a system that computes quantitative values of trust for the text in Wikipedia articles by visualizing the trust of a word in an article, which is computed on the basis of the reputation of the original author of the word. Suh et al. [26] introduced a social dynamic analysis tool, WikiDashboard to improve social transparency by surfacing hidden social context of pages and articles of Wikipedia.

4. Method

Having described the details of how Wikipedia works currently, we now turn our attention to modeling the growth model. We aim to analyze the growth of Wikipedia and various editing activities. We downloaded and analyzed a database dump of English Wikipedia offered from http://download.wikipedia.org. The dump file used in our study was generated on Oct 8, 2008 and contains the metadata of all edits (224,473,632 revision) made on English Wikipedia pages (14,915,993 pages). Each edit record of the dump file contains detail information of when (timestamp), who (user id), and where (page id) the edit was made, as well as the revision comment, which is an optional textual summary of the edit.

We acquired various Wikipedians' activities (e.g., regular edit, reverted edit, making revert, vandalism) by analyzing the edit records. The wiki platform underlying Wikipedia, MediaWiki, uses a single data format to store most of the activities. To distinguish different types of activities, we examined *revision comment* and *user id* of the edit records. To process the data, we utilized Hadoop [12] and Pig [21], as part of a distributed software platform for storing and processing a large-scale data.

The activities of *bots* were acquired by collecting edits made by a user whose user id is registered as a bot. The same technique was used to collect the activities of administrators. The activities by both types of actors are summarized in Figure 2.

As discussed in section 3.1, *revert*, *vandalism*, and *counter-vandalism* are notable activities in Wikipedia. These activities were captured by examining *revision comment* of edits. When the revision comment of an edit contains "vand", "spam", or "rvv" (Wikipedians' acronym of *revert due to vandalism*), the edit was recognized as a counter-vandalism activity and its immediately preceding edit was marked as vandalism.

Revert activities were extracted in a similarly way. We examined *revision comment* to see if it contains "rv", "revert", or "undid" to recognize revert. Note that this method cannot capture all relevant activities since it relies on optional textual summary annotation, added by the editor. For example, when an editor make reverts an edit without leaving any *revision comments*, it is not possible to apply this method. However, our earlier study [15] suggested that this method is a good approximation for identifying reverts. We believe that it is a reasonable method for comparison studies such as analyzing trends.

In the results presented in this paper, vandalism and *bot* edits are generally excluded in the analyses. Vandalism does not add any value to Wikipedia and dealing with the vandalism (e.g. repairing content in response to the vandalism edit) can be regarded as pure maintenance overhead. Edits by *bots* also does not involve human activity since they are launched to do repetitive jobs. In this paper, we focus on Wikipedians' activity as knowledge generation and its evolution.

5. Results

In this section, we present and depict the results of our analysis in the changes of global activity, details of the activities by editor classes, the evolution of editor population, and some analysis of the resistance they experience.

5.1 Slowing Growth in Global Activity

We first show that there is an overall slowdown in global editing activity in Wikipedia. Earlier studies [5][15][29][30][37] showed that Wikipedia has been grown exponentially and the growth follows the "edit begets edit" model or power law [30] as we discussed in earlier sections.

We analyzed the Wikipedia edit log to investigate recent growth pattern. Our analysis on the recent data set shows a strikingly different picture from what was reported two years ago. As shown in Figure 3, the global edit activity has stopped growing since early 2007.

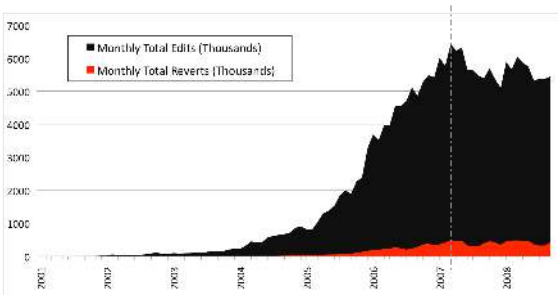

Figure 3: Monthly edits and identified revert activity.

The upper curve in Figure 3 represents monthly edit activity measured by the number of edits made in a month. The bottom curve in Figure 3 represents monthly revert activity. It is clear to see that reverts has been kept to less than 6% of total edit activity. However it has been increasing from 2.9% in 2005, 4.2% in 2006, 4.9% in 2007, and finally to 5.8% in 2008.

Not only has the total monthly editing activity stopped growing (as shown in Figure 3), but also the total number of active editors each month stopped increasing in 2007. Figure 4 shows the trend of Wikipedia editor population. Since its peak in March 2007 (820,532), the number of monthly active editors in Wikipedia has been fluctuating between 650,000 and 810,000. This finding suggests that the conclusion in [2][24] may not be valid anymore.

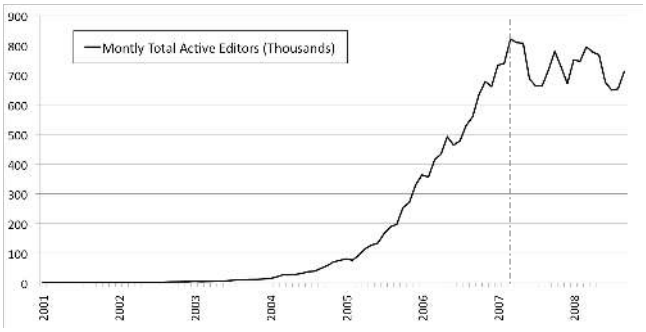

Figure 4: Monthly active editor - number of users who have edited at least once in that month

Similarly, Wikipedians themselves analyzed the growth of articles in Wikipedia. The Wikipedia community, who initially adopted an exponential growth model [36], developed a logistic growth model as shown in Figure 5.

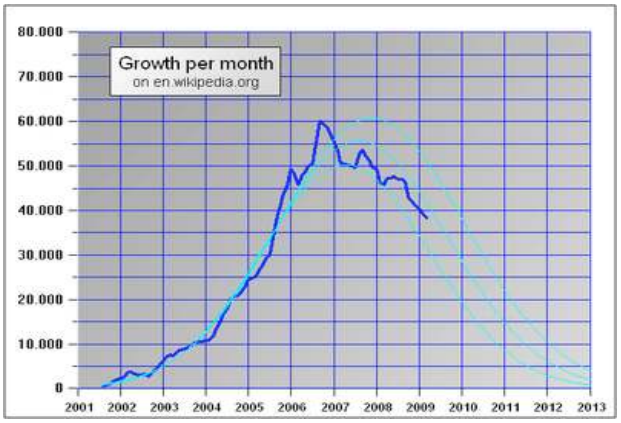

Figure 5: Article growth per month in Wikipedia[3]. Smoothed curves are growth rate predicted by logistic growth bounded at a maximum of 3, 3.5, and 4 million articles.

Figure 5 shows that article growth reached a peak in 2007-2008 and has been on the decline since then. This result is consistent with a growth processes that hits a constraint – for instance, due to resource limitations in biological systems. Microbes grown in culture will eventually stop duplicating when nutrients run out. Rather than exponential growth, such systems display logistic growth as in Figure 6.

[3] http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth

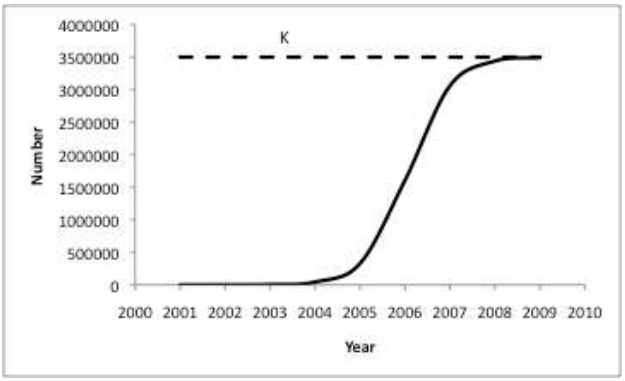

Figure 6: A hypothetical logistic Lotka-Volterra population growth model bounded by a limit K.

Figure 6 was generated by a Lotka-Volterra population model that assumes a resource limitation K (this is meant purely for illustration).[4] At the early stages of population growth the growth rate appears exponential, but the rate decelerates as it approaches the limit K. If the total amount of encyclopedic knowledge were some constant K, then the write-up of that knowledge into Wikipedia might be expected to follow a logistic such as Figure 6.

But there is a general sense that the stock of knowledge in the world is also growing. For instance, studies of scientific knowledge (e.g., [13][23]) suggest that it exhibits exponential growth. Also, events in the world (e.g., the election of Barack Obama or Lindsey Lohan's rehabilitations) create new possibilities for write-up.

However, even if the total amount of knowledge exhibited some monotonic growth as a function of time, K(t), one might still expect a variant of logistic growth as depicted in Figure 7.

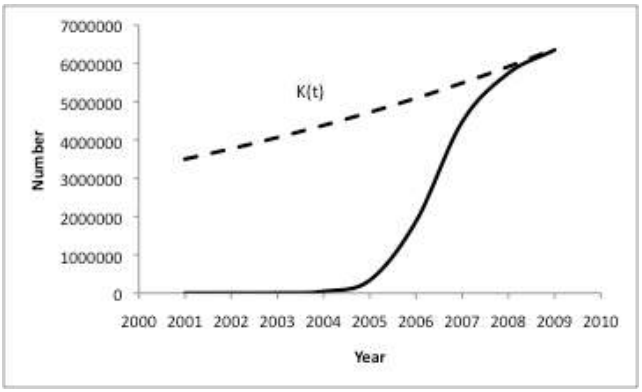

Figure 7: A hypothetical Lotka-Volterra population growth model bound by a limit K(t) that itself grows as a function of time.

As originally recognized by Darwin in relation to the growth of biological systems [7], competition (the "struggle for existence") increases as populations hit the limits of the ecology, and advantages go to members of the population that have competitive dominance over others. By analogy, we suggest that (a) that the population of Wikipedia editors is exhibiting a slowdown in its growth due to limited opportunities to make novel contributions, and (b) the consequences of these (increasing) limitations in opportunities will manifest itself in increased patterns of conflict and dominance.

The limitations in opportunities might be the result of multiple and diverse constraints. For example, on one hand, we expect that the capacity parameter K is determined by limits that are internal to the Wikipedia community such as the number of available volunteers that can be coordinated together, physical hours that the editors can spend, and the level of their motivation for contributing and/or coordinating. On the other hand, we expect that the capacity depends also on external factors such as the amount of public knowledge (available and relevant) that editors can easily forage and report on (e.g., content that are searchable on the web) and the properties of the tools that the editors and administrators are using (e.g., usability and functionalities). See the discussion section for more details.

In summary, globally, the number of active editors and the number of edits, both measured monthly, has stopped growing since the beginning of 2007. Moreover, the evidence suggests they follow a logistic growth function.

5.2 Analysis of Activity by Editor Classes

This section characterizes in more details the slowdown in growth of Wikipedia activity, specifically around different editor classes. For each month, we first partition the editors into different classes based on their monthly editing frequency. We then compare the total edit activities among the different editor classes over time.

In order to partition the editors into different classes, we first need to understand the edit activity distribution amongst the editors. In Wikipedia, the population of editors follows a power law distribution (also known as the long tail distribution) [30]. That is, relatively few highly prolific users account for a large percentage of the overall editing activity. A large population (the long tail) of less prolific editors contribute the rest of the content [16].

Consistently with the power law, we classified users using an exponential scale: we defined the classes of editors using powers of 10, e.g. 10^0, 10^1, 10^2. This resulted in five classes of users for each month: editors contributing 1 edit (i.e., 10^0), 2 to 9 edits (2-9 class), 10 to 99 (10-99 class), 100 to 999 (100-999 class), and more that 1000 edits (1000+ class). For example, we expected that the editors who contribute 1 edit only (monthly) would behave differently from others. Note that the classification of the editors was recalculated for each month.

By breaking down the global edit activities, we can now analyze the growth trends of the total monthly edits from the five editor classes. Figure 8 shows the number of edits contributed monthly by the five editor classes defined above. Since the beginning of 2007, the trends of four classes slightly decrease their monthly edits. In contrast, only the highest-frequency class of editors (1000+ edits, dark blue line) shows an increase in their monthly edits.

[4] http://en.wikipedia.org/wiki/Logistic_function

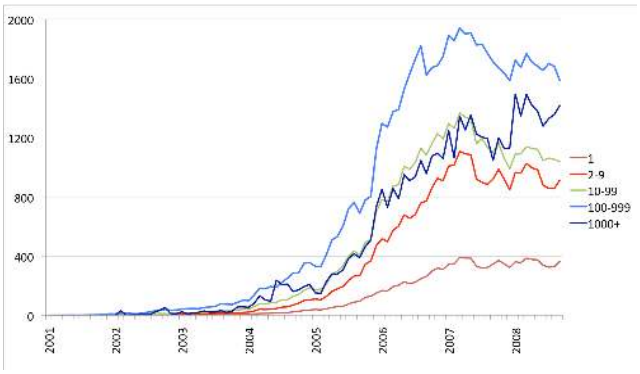

Figure 8: Monthly edits by user class (in thousands).

Another way to look at this data is to analyze the relative amount of activities for each editor class by transforming the data into percentages of the total edits. Figure 9 complements the information in Figure 8 by showing the percentage of the volume of edits that each class contributes in relation to the total.

In Figure 9, the two highest frequency classes of editors account for more than half of the total monthly edits (56% from 01/2005 to 08/2008). Furthermore, since 2005 the proportion of contributions by the highest-frequency editor class has increased slightly. In fact, the editors in 1000+ class have kept producing at an increasing rate over the past four years (their average monthly edits per editor for the years 2005 to 2008 were 1740, 1859, 1869, and 2095, respectively).

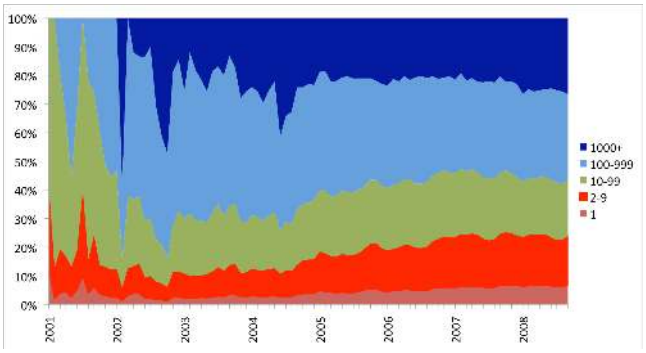

Figure 9: Monthly percentage of edits by each user class.

The results presented above illustrate the slowdown in contributions. We now focus on specific evidence about what might have contributed to such slowdown. *Revert* is the action of deleting a prior edit. Figure 10 shows the percentage of edits that were reverted (*reverted edits*) monthly for each editor class. Note that edits related to vandalism and edits performed by robots are excluded.

Figure 10 illustrates two indicators of a growing resistance from the Wikipedia community to new content. First, in contrast to the global slowdown that we saw in the previous section, the figure shows that the total percentage of edits reverted increased steadily over the years, without any slowdown. The total percentage of monthly reverted edits (see dashed black line in Figure 10) has steadily increased over the years for the all classes of editors (e.g. 2.9, 4.2, 4.9, and 5.8 percent of all edits for 2005 through 2008 as shown by the dash line in Figure 10).

Second, more interestingly, low-frequency or occasional editors experience a visibly greater resistance compared to high-frequency editors. Since 2003, edits from occasional editors have been reverted in a higher rate than edits from prolific editors. Furthermore, this disparity of treatment of new edits from editors of different classes has been widening steadily over the years at the expense of low-frequency editors. We consider this as evidence of growing resistance from the Wikipedia community to new content, especially when the edits come from occasional editors.

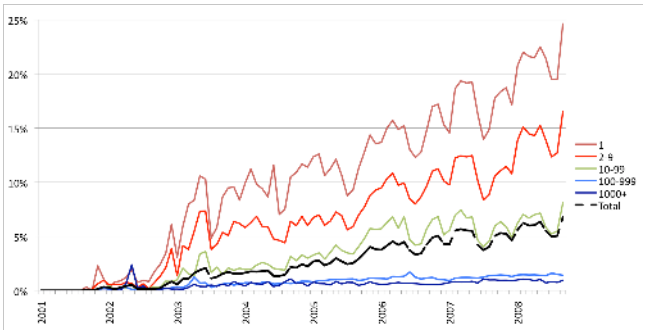

Figure 10: Monthly ratio of reverted edits by editor class

In the next section, we relate this observation to the population of editors that are active monthly for each class.

5.3 Analysis of Population by Editor Class

To investigate which factors affected the slowdown in edit growth, we examine the evolution of the population of active editors. The stalled growth of edit activity described in the above sections might be the result of two explanations: fewer active editors or fewer edits per editor.

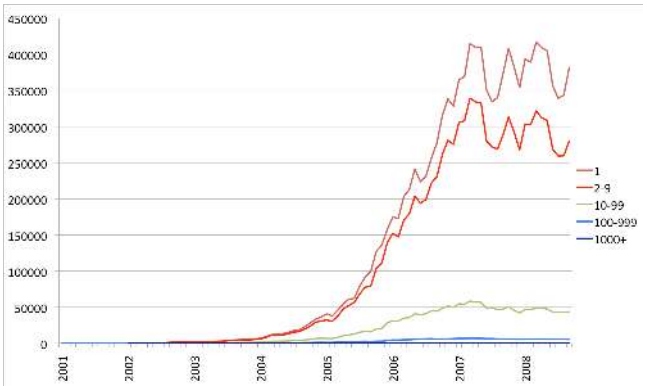

Figure 11: Monthly active editors by editor class. (This is a breakdown of the total editor population depicted in Figure 4)

We use the same editor classification as previous sections to count the number of active editors in each month. Figure 11, Figure 12, and Figure 13 show three views of the evolution of the population of the five editor classes.

Figure 11 shows the monthly frequencies of active editors by class. As expected from the power law distribution [30], the distribution of editors is very skewed: most of the editors contribute very few edits and very few editors contribute most of the edits. In fact, the two most prolific classes of editors (100-999 and 1000+) account for only about 1% of the population, but they contribute about 55% of edits (33% and 23% respectively).

Figure 12 uses a logarithmic scale to show the consistent slowdown of the growth among all editor classes over time, which

is not clear in Figure 11 for editors in 100-999 and 1000+ classes. The monthly population of active editors stops growing after March 2007: a surprisingly abrupt change in the evolution of the Wikipedia population for all the editor classes. This change is consistent with the slowdown of the editing activity shown in Figure 3.

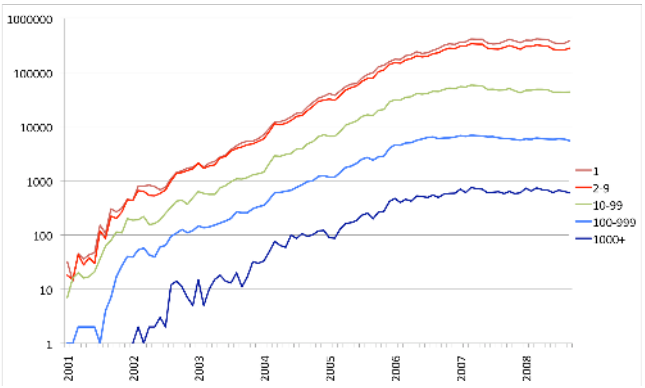

Figure 12: Monthly active editors by user class. The vertical axis uses a logarithmic scale.

Figure 13 shows the percentage of monthly active editors among the five classes. Note that the Y-axis is truncated: it omits the bottom 50% which represents the very long tail of once-monthly-editors.

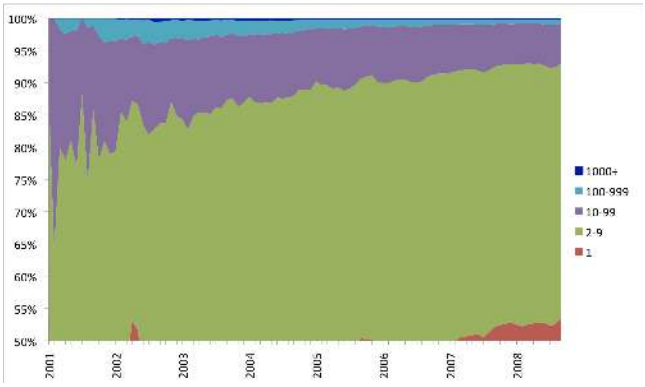

Figure 13: Contribution ratio of monthly active editors by their class. Note that the graph is truncated to highlight the declining population of 2-9 and 10-99 editor classes.

After March 2007, the populations of the five editor classes show different trends of evolution: these are shown in terms of absolute numbers in Figures 11 and 12 and in terms of percentages of the total population of active editors in Figure 13. The population of the highest-frequency editors (1000+ class) levels off (average monthly editors in 2005, 2006, 2007, 2008 were 204, 505, 643, 665, respectively) while the populations of the two intermediate classes (10-99 and 100-999 class) exhibit a descending trend: from 9% of 10-99 editors and 1.3% of 100-999 editors in 2005 to, respectively, 6% (10-99 editors) and 0.8% (100-999 editors) in 2008. The editors in the 2-9 class remains relatively stable (41% in 2005 and 40% in 2008). Finally, the editors in the lowest-frequency class (1 edit-per-month) increase slightly.

In summary, the results show that while the subpopulations of the highest-frequency (1000+) and lowest-frequency (1 edit-per-month) classes of editors stop growing and then remain about stable, the middle frequency classes gradually decrease their proportion. The differences among the classes over time suggest that editors in mid-frequency are moving toward a lower frequency class, resulting in a shrinking middle class.

We see these trends as being closely related to the growing resistance toward occasional editors illustrated in the prior section. A closer look at the most recent data from 2008 suggest that these trends might continue in the near future, such as the reduction in the population of the middle class. Further research will be required to see if the same trend will also affect with some delay the class of top editors.

5.4 Block, Protection, and Page Delete

In this section, we present analyses of other types of activities performed by Wikipedians that suggest increasing resistance and protectionism.

As discussed in section 3, editing is the primary activity in Wikipedia. However there are also other types of activities. For example, "page deletion" is not recorded as an edit, but is logged as a different type of action in MediaWiki.

User Block is an administrative action that prevents a user from making contributions. Blocks are sometimes used as a deterrent, to discourage certain behavior and encourage a productive environment. Vandalism is a common cause of blocking, as enforced by administrators.

Figure 14 compares patterns of *blocking* and *vandalism*. The average ratio between reported vandalism edits and blocking of editors is about 5:1. It is interesting to note that vandalism edits and blocking have similar trends. In other words, the blocking of an editor seems to follows reports of vandalism activity. This suggests that *blocks*, policing actions enforced by administrators, might be an effective tactic for deterring vandalism.

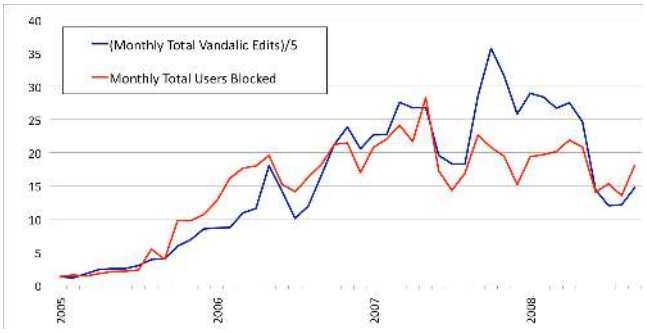

Figure 14: Monthly block (in thousands) and vandalism (in units of 200). Vandalism is scaled down by a factor of 5 to show the correlation.

Page Protection is an action in which an administrator protects a page from being edited or moved. Such protection might be indefinite, or might expire after a specified time. Protection is typically used when there is a content dispute or sustained vandalism. Figure 15 shows monthly new page protections over time along with monthly vandalism for comparison. Compared to User Blocks, the relationship between protection and vandalism is not as clear. However, the number of protection incidents is increasing, which indicates more administrative actions are being imposed by the Wikipedia community.

Note that the number of page protections presented in the figure is the number of newly enforced protections rather than the total number of pages under protection (i.e., the same page can be

protected and unprotected multiple times). The ratio of protected pages over the total number of pages would be a reasonable measure to assess the degree of protection in Wikipedia. However, we found that it is technically challenging to collect all the required information. Further research is required to investigate how *protection* has been affecting the social dynamics in Wikipedia.

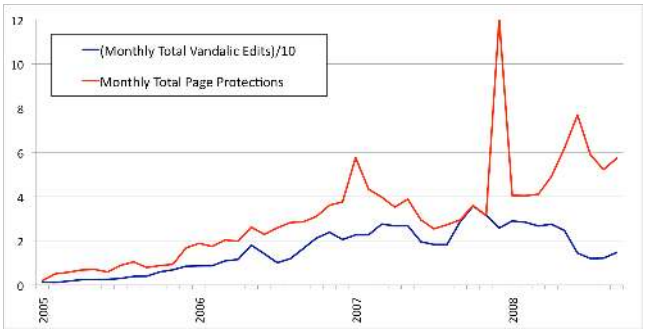

Figure 15: Monthly *new* page protection (in thousands) and vandalism (in hundreds)

Page Delete is an action in which a Wikipedia administrator removes the current version and all previous versions of an article from public view. Reasons for deletion include, but are not limited to; copyright violation, vandalism, and articles whose subject fails to meet the "notability" threshold. Generally, content regarded as not suitable for an encyclopedia is deleted. However, there is a grey area to this decision and the Wikipedia community has argued about what should be included in Wikipedia [18]. Deleted pages can be restored if necessary, which is also logged.

Figure 16 shows the page deletion activity of Wikipedia since December 2004 along with the number of pages created. The number of *page deleted* activity was extracted from the Wikipedia *logging* table while the number of *page created* activity was calculated by counting the number of first edits made to any pages. We also collected data about the *page restore* activity. However, we found that only 1.8% of the *pages deleted* were restored later. Due to its low frequency, the number of *page restore* operations is not included in the graph.

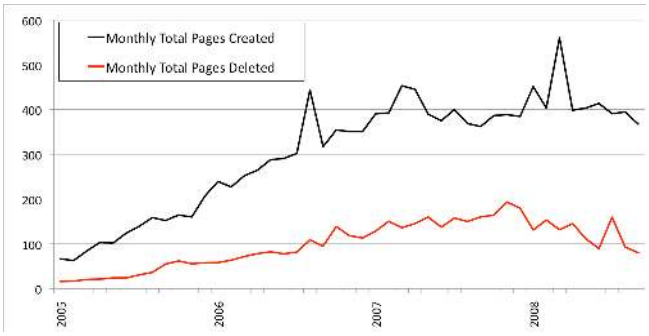

Figure 16: Monthly page creation recorded (page survived) and page deleted (both shown in thousands)

On average, since January 2005, 25% of the pages created are deleted suggesting 3 out of 4 pages created survived. However, the survival rate of newly created page has been decreasing as shown in Figure 16. During 2005, 34,889 pages were deleted while 126,904 pages survived (78%). The survival rate kept deteriorating to 72% (394,640 page survived, 155,174 page deleted) by the end of 2007. We view this as another indicator of growing resistance to new content from the community.

Interestingly, the data shows an increasing trend of pages being restored during 2008 and the survival rate of new pages remains stable at a slightly increased level. We do not have an explanation for this new trend in 2008 and further research is required to explain this new pattern.

On the other hand, at the end of 2005, Wikipedia changed its editorial policy so that anonymous users cannot create an article anymore. It is interesting to see that this policy change did not affect the increasing number of *page deletes*. This would suggest that page deletions are part of the regular maintenance for the entire community rather than a specific response to anonymous users.

In performing this analysis, we assumed that the deletion of a page is determined within a month after the page is created. Then, when a page has survived for a month, we assume that it is less likely to be deleted afterward. This assumption was based on qualitative observations of the Wikipedia editorial practices.

In summary, together with the increasing rate of reverts shown above, the analyses of various administrative activities provide additional indicators suggesting an increasing tendency to resist new contributions.

6. DISCUSSION

Wikipedia can be viewed as an in-vivo, unprecedented experiment on knowledge co-creation. It is in-vivo because, together with search engines, it is a real case of collective intelligence that has changed how people typically search and consume encyclopedic knowledge compared to a few years ago [20]; in turn, this is also affecting the design of traditional encyclopedias [10]. It is an unprecedented experiment because of the community (i.e., its scale and bottom-up self-organization), the amount of knowledge shared, and the web tools used by editors and administrators to share and manage the entire process. These properties make Wikipedia a very appealing object of investigation for social and technology scientists. However, due to the lack of prior comparable socio-technical systems, researchers do not have sound theoretical models to make robust predictions about the evolution of Wikipedia. This dearth of models motivated our line of research.

Studying the evolution of the activity and population of editors over time has informed us that prior exponential growth models for Wikipedia need to be critically re-visited. We found that Wikipedia has stopped growing over the last two years.

Among various factors, our study suggests that the following may have affected the growth of Wikipedia:

(a) the growing resistance to new content especially when contributed by occasional editors;

(b) the greater overhead imposed by the costs for coordination and bureaucracy;

(c) the availability of easy topics to write about;

(d) the quality of the tools used by editors and administrators.

Growing resistance: We found a number of indicators of growing resistance. The rate of reverts-per-edits (or new contributions rejected) and the number of pages protected has kept increasing. Occasional editors experience a greater percentage of reverts per edits in comparison to the more prolific editors. The total number

of *blocks* and *page deleted* increased. Restrictive policies have been introduced. For example, since December 2005 anonymous users could not create new articles and articles could be protected from changes by new users.

The greater resistance towards new content has made it more costly for editors, especially occasional editors, to make contributions. We argue that this resistance may have contributed, with other factors, to the slowdown in the growth of Wikipedia. These data appear consistent with the hypothesis that the "*deletionists*" may be increasingly outnumbering the "*inclusionists*" among the administrators [18].

Greater overhead: The amount of work devoted to coordination (e.g., maintenance and discussion) or community bureaucracy (e.g., formulating and discussing policies) in Wikipedia has increased as overhead costs to the community. Our work in the past [15] found that the costs due to coordination have grown visibly from 2001 to 2006; as the proportion of work devoted to coordination increased, the proportion of work devoted to content production decreased. Our data about *administrators'* and *bots'* edits appear consistent with these prior findings. Bureaucracy in Wikipedia has also increased as the community has grown in maturity and importance, as a social system [4].

These types of overhead have a negative impact on content production because (a) they take efforts away from the direct work on content and (b) it may affect the motivation of the contributors. For example, administrators may be less excited to perform housekeeping. Newcomers may more easily get lost in a strict bureaucracy[18]. Those findings appear consistent with prior research about the mechanisms that regulate communities. For example, Benkler [3] points to the factors that regulate coordination and information flow in online peer-production. Others point to the factors that direct self-organization and self-governance in human communities [25][28].

Finally we consider two external factors that might play a role in constraining the growth of Wikipedia activities and editors. We have not directly measured them but we speculate about their possible role in the slowdown.

Running out of easy topics: Wikipedians might have already taken care of the "low-hanging fruit", having compiled articles on common topics. 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. As the number of such easy topics gradually diminishes, the competition on the remaining few increase. Alternatively, the work on harder topics requires editors to invest greater time and effort. As suggested by Figure 6, even though there are new knowledge and events providing opportunities for Wikipedia entries, the space for non-specialist to make contributions diminishes significantly.

Editors' and administrators' tools. The factors listed above pertain mainly to constraints related to the editors or their activities (i.e., people-ware) and the external world's knowledge (i.e., knowledge-ware). In addition to them, we believe that the quality of tools may play an important role (i.e., tools-ware).

The Wikipedia project is novel for its world-scale and its bottom-up formation of community. As with architects who construct buildings of unprecedented scale, this community requires special infrastructure to build and maintain such a large shared encyclopedia. While policies, rules, and processes constitute the softer part of the infrastructure, the tools constitute the harder part, and enables production, sharing, and coordination. The cost of performing these actions depends on the usability and utility available to the administrators and editors. The greater the cost, the more difficult it is for Wikipedia to recruit and retain the best administrators and devoted editors.

Some limitations of our method will be addressed in future work.

We chose the month as our unit of analysis to aggregate the logs about editors or their actions. Alternative units of time can be used to analyze the evolution of seasonal, weekly, or daily patterns in Wikipedians' activities.

Also, we used semantically arbitrary boundaries when defining classes of editors. For example, we defined one of editor classes as the class of editors contributing between 10 and 99 edits monthly. More refined classifications could lead to more fine-grained distinctions among the editors.

In this paper, the editor classes presented in the graphs are recalculated monthly. Therefore, the editors' seniority (usually measured by how long a member is in a community) is not considered in our analysis. However, we believe the editors' seniority might have close relationship with their editing behavior.

Finally, our analysis was performed only with edits on the pages that remain publically available. When a page is deleted in Wikipedia, all the earlier edits made on the page are also removed from public view. The data that we used in this study does not contain those edits that had been removed.

7. Conclusion

Throughout this paper, we presented a number of new patterns in the evolution of Wikipedia.

The slowing growth of Wikipedia: We found a global slowdown of growth rate both in the number of editors and the number of edits per month. In addition, we analyzed editors' behavior in more detail by classifying them into editor classes based on their monthly edit frequency. The middle class of editors now cover a lower percentage of the total population, while high frequency editors continue to increase the number of their edits.

Greater resistance to new edits: The resistance, measured as the ratio of edits that are reverted, has steadily increased over the years for the entire community of editors (e.g. 2.9% in 2005 to 6% in 2008). More interestingly, occasional editors experience a greater resistance compared to high-frequency editors. Furthermore, the disparity of treatment by editor class has been widening over the years. Indicators such as page protection, page deletion, block, and other restrictive policies also display the same trend of increasing resistance.

This evidence is consistent with the interpretation that the growth is limited by available resources in Wikipedia, and advantages go to members of the population that have competitive dominance over others. In other words, Wikipedia is experiencing the evolution just like biology systems do and resulting in: (a) the slowing growth of the editor population due to limited opportunities in making novel contributions; and (b) increased patterns of conflict and dominance due to the consequences of the increasingly limited opportunities.

Our conclusions here are consistent with earlier studies, showing that a greater proportion of *the overall edits is being devoted to overhead activities* such as coordination, policy setting, and governance [4][9][15]. Our hope is that these findings are informative to designers of large-scale knowledge management

systems. Further research is needed to explore how these findings generalize to the evolution of other collaborative knowledge systems.

8. Acknowledgements

We would like to thank Robert Rohde (UC Berkeley) and Andrew Lih for helpful discussions.

9. REFERENCES

[1] Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., Raman, V. Assigning Trust to Wikipedia Content. In WikiSym 2008: International Symposium on Wikis, 2008

[2] Almeida, R.B.m, Mozafari, B., and Cho, J., On the evolution of Wikipedia. ICWSM 2007, Boulder, Co., 2007.

[3] Benkler, Y. Coase's Penguin, or, Linux and The Nature of the Firm. The Yale Law Journal, Vol 12, N 3. December 2002.

[4] Butler, B.S., Joyce, E., and Pike, J. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. CHI 2008, 1101-1110, 2008.

[5] Capocci, A., Servedio, V., Colaiori, F., Buriol, L., Donato, D., Leonardi, S., and Caldarelli, G. Preferential attachment in the growth of social networks: the Internet encyclopedia Wikipedia. PRE, 74(3):036116, 2006.

[6] Chi, E.H. A position paper on 'Living Laboratories': Rethinking Ecological Designs and Experimentation in Human-Computer Interaction. Presented at 2009 HCIC Workshop. Frasier, Colorado, Feb 2009.

[7] Darwin, Charles (1859), On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (1st ed.), London: John Murray

[8] Denning, P. Horning, J., Parnas, D. and Weinstein, L. Wikipedia Risks, CACM 48(12), Dec. 2005.

[9] Forte, A. and Bruckman, A. Scaling Consensus: Increasing Decentralization in Wikipedia Governance. Proceedings of HICSS-41, January 7-10, 2008.

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

[11] Gordon E. "Cramming more components onto integrated circuits". Electronics Magazine, 4, 1965.

[12] Hadoop, http://hadoop.apache.org/core/

[13] He, X. and Zhang, J. On the Growth of Scientific Knowledge: Yeast Biology as a Case Study. PLoS Comput Biol 5(3), 2009

[14] Kamm, O. Wisdom? More like dumbness of the crowd, The Times, Aug 17 2007. http://www.timesonline.co.uk/tol/comment/columnists/guest_contributors/article2267665.ece

[15] Kittur, A., Suh, B., Pendleton, P.A., and Chi, E.H. He Says, She Says: Conflict and Coordination in Wikipedia. In Proc. of ACM Conference on Human Factors in Computing Systems (CHI2007), 453-462, April 2007.

[16] Kittur, A., Chi, E., Pendleton, B.A., Suh, B., and Mytkowicz, T. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In Alt.CHI, 2007.

[17] Kurzweil, R. The Singularity Is Near, Viking Penguins, 2005.

[18] Lih, A. (interview), The Battle for Wikipedia's Soul, The Economist (magazine), March 6th 2008.

[19] Lyman, P. and Varian, H.R., "How Much Information", 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on April 1, 2009

[20] Malone, T. W., Laubacher, R., and Dellarocas, C. N. Harnessing Crowds: Mapping the Genome of Collective Intelligence. MIT Sloan Research Paper No. 4732-09, 2009

[21] Pig, http://hadoop.apache.org/pig/

[22] Priedhorsky, R., Chen, J., Lam, S.K., Panciera, K., Terveen, L., and Riedl, J. Creating, Destroying, and Restoring Value in Wikipedia, In Proc GROUP 07, 259-268, 2007

[23] Price, D.J.d.S. Little science; big science. Columbia University Press, New York, 1963.

[24] Spinellis, D., and Panagiotis, L. The collaborative organizations of knowledge. Communications of the ACM, 51(8), 68-73, 2008.

[25] Stvilia, B., Twidale, M., Smith, L. C., and Gasser, L. Information quality work organization in Wikipedia. JASIST, 59(6), 983–1001, 2008.

[26] Suh, B., Chi, E.H., Kittur, A., and Pendleton, B.A. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. Proc. CHI '08, 1037–1040, ACM Press, 2008.

[27] Viégas, F.B., Wattenberg, M., and Dave, K. Studying cooperation and conflict between authors with history flow visualizations. Proc. CHI '04, 575–582. ACM Press, 2004.

[28] Viégas, F.B., Wattenberg, M., Mckeon, M. The Hidden Order of Wikipedia, Online Communities and Social Computing, pp. 445-454, 2007

[29] Voss, J.. Measuring Wikipedia. In Proc. ISSI 2005, Stockholm, 2005.

[30] Wilkinson, D. and Huberman, B. Assessing the value of cooperation in Wikipedia. First Monday, 12(4), 2007.

[31] Wikipedia, Administrators, http://en.wikipedia.org/wiki/Wikipedia_administrator

[32] Wikipedia, Assessment http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

[33] Wikipedia, Essjay controversy http://en.wikipedia.org/wiki/Essjay_controversy

[34] Wikipedia, Reliability of Wikipedia, http://en.wikipedia.org/wiki/Reliability_of_Wikipedia

[35] Wikipedia, Wikipedia http://en.wikipedia.org/wiki/Wikipedia

[36] Wikipedia, Wikipedia Statistics, http://en.wikipedia.org/wiki/Wikipedia:Stats

[37] Zlatic, V., Bozicevic, M., Stefancic, H., and Domazet, M..Wikipedias: Collaborative web-based encyclopedias as complex networks. PRE, 74(1):016115, 2006.