# STATISTICAL METHODS IN ECOMMERCE RESEARCH

## Network Analysis of Wikipedia

**Robert H. Warren**
School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada

**Edoardo M. Airoldi**
Computer Science Department & Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544 USA

**David L. Banks**
Department of Statistics, Duke University, Durham, NC 27708, USA

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

## CHAPTER 1

# NETWORK ANALYSIS OF WIKIPEDIA

**Abstract:** This chapter analyses data on the growth and connection of Wikipedia, the on-line collaborative encyclopedia. The data suggest at least three stages of growth, the last of which has only recently emerged. We also consider how growth depends upon infrastructure and internal links.

## 1.1 INTRODUCTION

The Wikipedia is an on-line encyclopedia created by the volunteer efforts of Internet users all over the world. Although it is not a commercial enterprise, it has relevance to e-commerce activities. In particular, companies with business models in which value is created by users who link sites and share content should be interested in the dynamics of network growth seen in the Wikipedia data.

Some companies of this kind generate revenue through advertising, or provide such valuable functionality that their enterprise is bought by others. Others do not offer their network directly to the world, but use it internally to manage and annotate memos, files, and accounts. Snapfish, for photo sharing, is an example of a for-profit company that generates revenue chiefly through advertising. Facebook is an example of a social networking tool that has become so popular that Google thinks it will draw people to them, although their specific business plan is unclear. And John Negroponte, the U.S. intelligence czar, says his analysts use an in-house system called "Intellipedia" to manage their internal information sharing (Reuters, 2006). There are many other examples, and no doubt more are coming.

### 1.1.1 Motivation

This paper uses data on the initiation and editing of Wikipedia entries to understand growth, revision, and linkage in a complex, multi-user network.

Specifically, organizations considering the use of wiki's will want to weight the benefits of a wiki with the preparation required. Beyond hardware and software requirements, there is the question of the amount of initial "bootstrapping" content that the organization must provide in order to make the wiki usable and attractive to the target audience.

This initial content varies in its quantity and sophistication, and we wished to know the relationship between the initial content provided by the organization and the content that would be added by the user.

Additionally, most wiki's allow for the creation of category data that classifies and links the different content pages in a taxonomy. Since the difference between category and content is under continual debate in the expert fields, we thought it interesting to look at the category taxonomy constructed by the end users. Would they use the skeleton provided by the organization, provide their own categorical data, or ignore the categories altogether?

Finally, there exists a debate on whether wiki methods are a worthwhile new means of representing information, separate from the online journals (blogs)and commented lists-of-links (web-logs) that a prevalent on the Internet. If wiki's are expected to be online references, then they should be more than just lists of links and provide new and unique user-generated content.

Looking forward, we believe that e-commerce business enterprises that attempt to emulate Wikipedia's strategy for creating value will want to benchmark their own networking projects against the growth dynamics of Wikipedia; it is likely that the rapid growth of Wikipedia reflects a fortunate confluence of circumstances that deserves study and replication. Some of the key questions in this line include:

- What is the balance between the amount of effort in creating new entries compared to editing and correcting entries, and how does this change over time?

- Are there economies of scale in managing a collaborative project? Does the growth of Wikipedia suggest a cartoon model with typical phases of growth?

Answering these questions will provide insight into the growth of a major new Internet phenomenon, and perhaps guidance to those who attempt to mimic its success.

As a caveat, one of the challenges in this kind of research is that the best sources are on-line, and Wikipedia itself maintains some of the most useful information and analyses. But these sites can be revised at any time; in particular, posted information may be updated or removed. The material used in this paper is current as of May 8, 2007. A related problem, much less important but aesthetically irksome, is the fact that the use of URLs in citation causes wordprocessing systems to balk, producing line overruns or introducing potentially misleading dashes. The world needs a convention for hyphenating URLs; we propose and use the $\oplus$ sign, since that cannot appear as a character in any URL and thus avoids ambiguity.

## 1.2 BACKGROUND ON WIKIPEDIA

There are many Wikipedias, divided according to language and largely independent of one another. This article focuses on the first and largest, which is the English Wikipedia. The (English) Wikipedia was conceived in 1999, the creation of Jimmy Wales and Larry Sanger as a project at Wales' company Bomis. Bomis built Nupedia, an on-line encyclopedia with free content, but the articles were produced by selected experts and refereed for content. The recruitment of experts and the refereeing of the articles led to substantial delays; thus Wales and Sanger decided to drop that entirely and create a software system that allowed volunteers to create and post articles, which could then be revised and improved by other volunteers. The term "wiki" is derived from the Hawaiian word for "quick" and presumably relates to the faster production time.

On January 15, 2001, Wikipedia went public. Its innovative approach had several immediate consequences. First, Wikipedia quickly spread beyond traditional encyclopedia topics to become a guide to popular culture—it contains articles on television shows, movies, and minor players on the world stage. Second, the accuracy of Wikipedia entries is less trustworthy than in hard copy encyclopedias (though not by much; cf. Giles, 2005). Third, there is the opportunity for mischief and vandalism. The comedian Stephen Colbert was banned from Wikipedia posting after he entered false information and encouraged his audience to do likewise (Pava, 2006). Also, the German version of Wikipedia was hacked so as to distribute copies of the Blaster worm (Leyden, 2006). Fourth, Wikipedia has become immensely popular; as of January 2007 it has grown to include more than 1.5 million articles in English, and about 5 million articles in about 250 languages (http://en.wikipedia.org/wiki/Wikipedia).

These kinds of issues and pressures meant that Wikipedia had to carefully track content creation, editing, and cross-linking. The key technology supporting wikis was developed by Ward Cunningham in 1995 (Leuf and Cunningham, 2001). The main feature of this technology is that each page has an "edit this topic" link that sends users to a control site that allows them to make changes to the topic. As part of that process, people can register with Wikipedia, creating a user profile (or, in about 25% of the cases, choose to remain anonymous). The version control system for Wikipedia tracks all changes to each topic, and which user (or IP address, for anonymous users) made those changes. This version control data is one of the main data sources for the analyses in this paper, along with the link structure for each topic and information from the user profiles.

The legal structure supporting the distribution of Wikipedia text is the GNU Free Documentation License. This permits anyone to use, modify, and distribute source code without limitation. That license has provision for "invariant sections" that cannot be changed except by the creator, even if they are inaccurate or plagiarized. Those portions of Wikipedia that are invariant pose content-management problems. But most content is not so governed, and contributors have no ownership; indeed, one of the facilitators of growth is the flexibility with which volunteers can correct and revise each other's work.

The thousands of Wikipedia volunteer editors collaborate to build a consensus on change. They review new entries, identify conflicts, and negotiate agreement on changes in content. The custom in the community is to avoid majority voting, although straw polls are used to get a sense of how the collective editorship stands. When disputes arise, as commonly happens, and no consensus emerges, the matter can be referred to a mediation committee; if that fails, Jimmy Wales has the authority to make the final decision. Individuals who gain prestige within the Wikipedia community through their discussion, mediation, and other contributions can obtain higher levels of privilege, such as the power to delete or freeze pages, or attain administrator status.

The development of this system of consensus building means that social networks among the editors play a large role in guiding the growth and content of the Wikipedia.

Section 2 describes the simple statistical features of Wikipedia growth and attempts to interpret those features in terms of a corporate growth model. Section 3 focuses upon the network features of Wikipedia, and how those have changed over time. Section 4 looks at the role of the social network within the Wikipedia community, and examines how that has affected growth. Section 5 draws general conclusions that may apply to similar efforts.

## 1.3 THE GROWTH OF THE ENGLISH WIKIPEDIA

Wikipedia is a fast-growing database. The success of the collaborative effort at the core of Wikipedia rests, in part, on its popularity as a source of information. Although such content does not go through the thorough vetting process that information in printed encyclopedias has to pass, the information available through Wikipedia has four crucial characteristics: it is quick to find, thanks to the many search engines that index its pages; it is good enough for a reader to get the big picture about an event, a person, or a difficult mathematical concept; it provides lots of useful pointers; and it evolves (by updating and self-correction) over time.

There are many metrics for growth, such as database size, the number of users, and the number of hits. Wikipedia itself is the primary source of data on its growth. Looking first at the number of megabytes over time, as available at `http://stats.wikimedia.org/EN/`, Figure 1.1 shows a classic exponential growth pattern.

The fitted model is $M = 11.39 \exp(T) - .00000383$, where $M$ is the number of megabytes and $T$ is time. The adjusted R-squared is .974, and the root mean squared error is 271.28; this indicates a very good fit. The natural interpretation is that the rate of growth of the Wikipedia is proportional to its size.

Figure 1.2 shows a similar pattern in the number of articles. The fitted model is $N(t) = N(0) \exp(t/\tau)$ where the estimated value is $\tau = 499.7$ and $N(0)$ is the starting value of 80 kilobytes on October, 2002. Note the small drop-off that occurs at the end of 2006, as the number of articles falls below the unsustainable growth of the exponential model.

From a commercial perspective, one of the most important metrics is "reach," an estimate of the number of people who use (or are specifically aware of) a product. Alexa Internet, Inc., is a company that captures statistics on site usage (through their toolbar product), and their data provide a picture of the growth of Wikipedia according to several different criteria (see `www.alexa.com` for more details). Figure 1.3 presents an estimate of the reach of Wikipedia from January 2003 through January 2006. The estimate is based on the the number of unique Alexa toolbar users who visit a site on a given day, with some smoothing over a rolling three-month period. (Thus the occasional downturn in the Fig. 1.3 does not represent people who forgot that Wikipedia exists, but rather a transient drop in the number

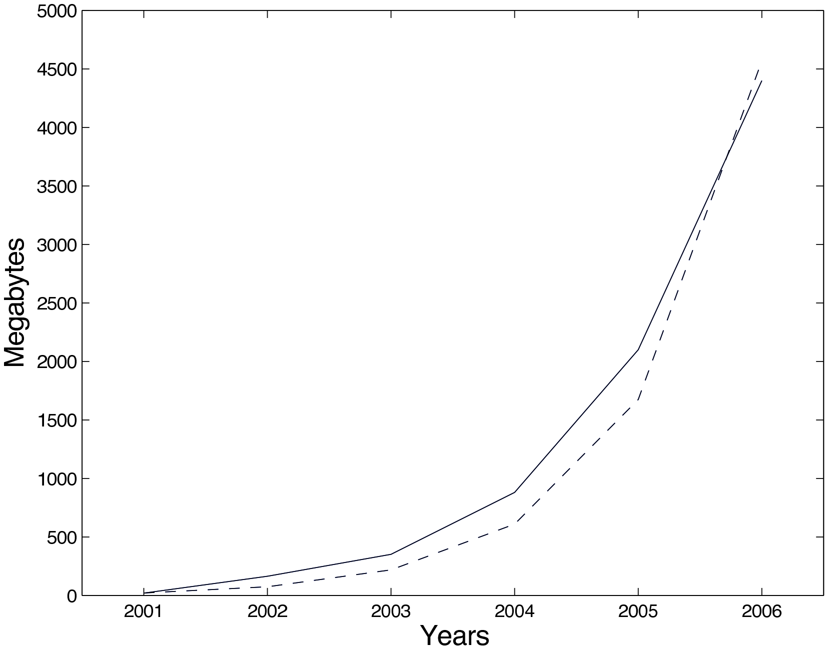

**Figure 1.1.** A plot of the number of megabytes in the English Wikipedia over time. The solid line represents the observed values, and the dashed line is a fitted exponential function.

of hits.) As of January 2006, Alexa estimated that one person in 50 knew about Wikipedia.

A second measure of success is traffic rank. Alexa's rankings are estimated from the proportion of Alexa toolbar users accessing top level domains, with a rolling three-month weighting scheme to smooth out short-term effects. Figure 1.4 presents the traffic rank of Wikipedia, as estimated by Alexa Internet, Inc., with a log scale for the rank axis. Note that on the log scale the growth trend is almost linear over this period, and that Wikipedia has rank about 30.

To provide context for the graphs in Fig. 1.3 and Fig. 1.4, Table 1.1 lists the top-ranked domains, in terms of traffic, for March 2, 2007. For these sites, the table also reports the percentages of population reached (a measure of awareness) and the percentage of pages viewed (a measure of active engagement), and the same statistics rescaled with Wikipedia's corresponding values—so that Wikipedia's

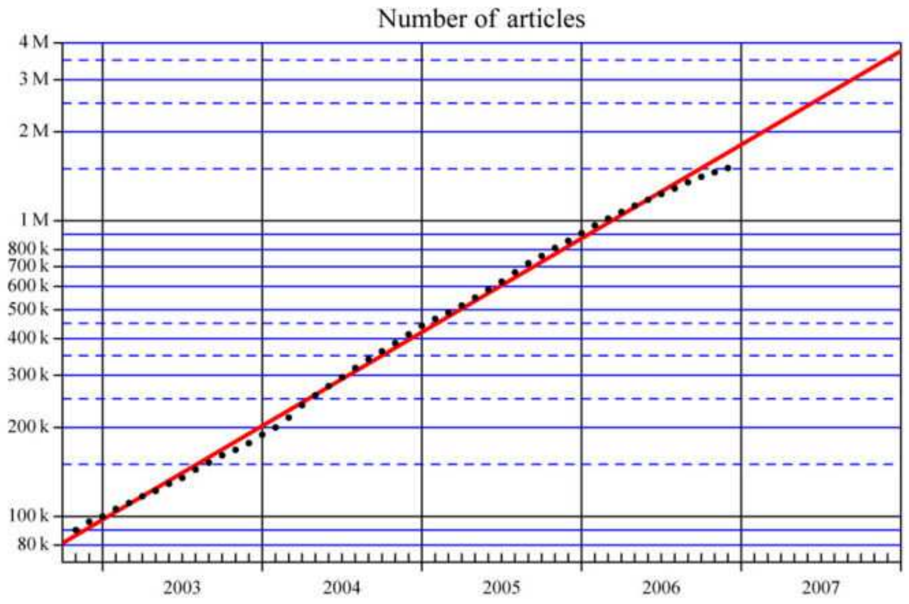

**Figure 1.2.** A plot of the number articles in the English Wikipedia over time. The number of articles is on the log-scale, and the line shows the best exponential fit. The image is from `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`.

R/W and V/W would equal one. Obviously, for an encyclopedia such as Wikipedia, the percentage of pages viewed will be relatively small on any given day. And note that Wikipedia's rank is 10, which accords well with the linear trend seen in Fig. 1.4 (that trend is clearly unsustainable, and the March 2007 data show some flattening, but it is clear that steady growth in rank has been a stable feature of the Wikipedia phenomenon).

With a nod to self-conscious post-modern reflexivity, Wikipedia collects some of its own traffic and usage data. These can be found at `http://en.wikipedia.o` ⊕ `rg/wiki/WP:AS`.

Reach and rank represent only two of many metrics for growth. Table 1.2 presents a selection of additional statistics that suggests how the database has evolved since its creation in 2001. In particular, the rapid increase in the number of non-English Wikipedias that contain more than 1000 articles (the last line in the table) is strong evidence of the popularity, portability, and perceived utility of the collaborative business model. As of March 2007 there are 242 non-English Wikipedia

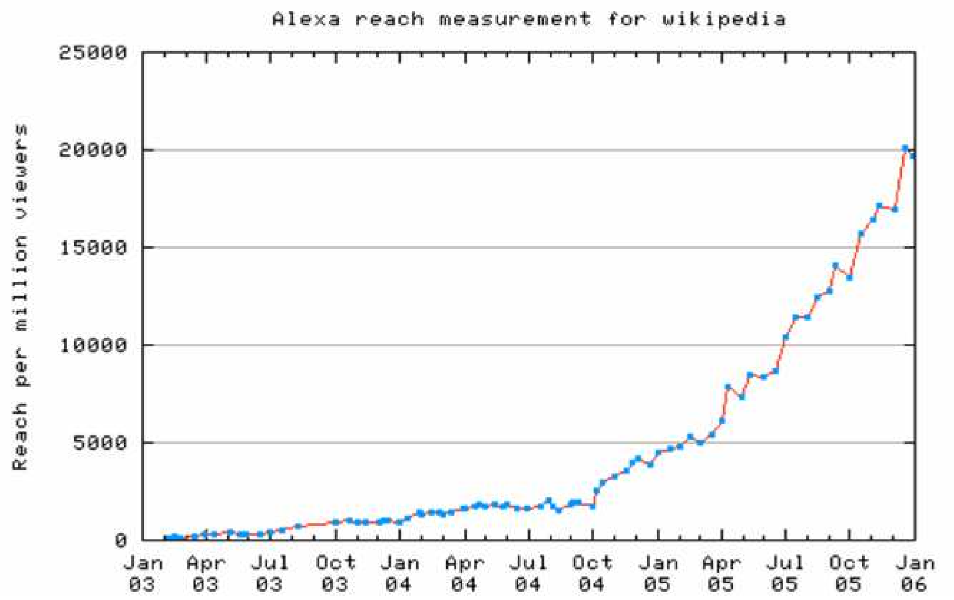

**Figure 1.3.** Reach per million measurement for wikipedia.org. (Source: Wikipedia/Alexa.)

sites that contain at least ten articles (http://en.wikipedia.org/wiki/Wik ⊕ ipedia:Multilingual_statistics).

As Table 1.2 suggests, a key aspect of the dynamics of the wikipedia growth concerns its structure. Images were introduced after a slight delay, in 2002. The introduction of categories, however, was even slower, and did not occur until May 2003. The late introduction of categories may be due to two tightly coupled sets of issues: on the one hand, it is good to wait and see what kind of content users contribute before putting management resources into defining a tree of labels (else many labels may remain unused). On the other hand, the size of the encyclopedia in its early stages may not have needed a formal tree of labels. But the number of categories has increased quickly over the last two years, from about 23,000 to about 176,000. Furthermore, the internal structure of the categories itself has evolved much over the years; compare, for example, the high level structure in November 2005 described by Holloway, Božičević, and Börner (2006) to the current one shown at http://stats.wikimedia.org/EN/CategoryOverview_EN_Concise.htm.

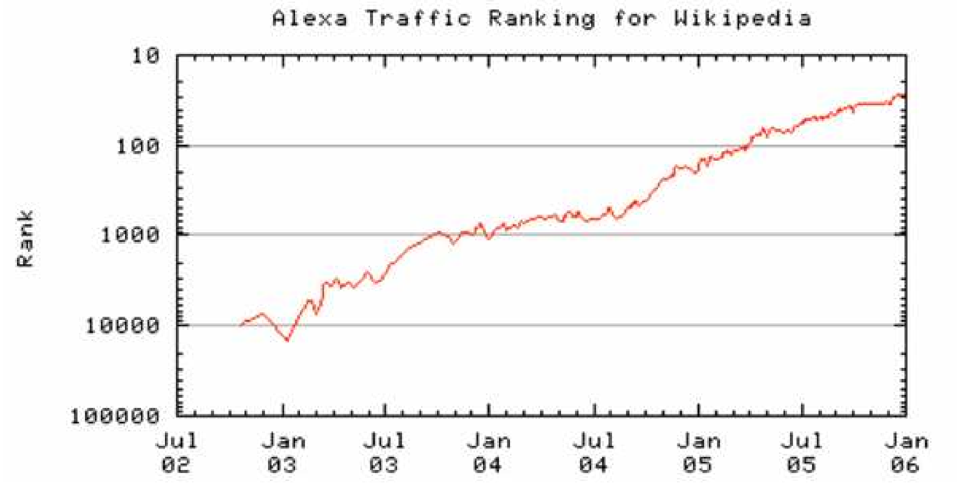

**Figure 1.4.** Traffic ranking for the English Wikipedia site. A rank equal to 1 would mean that Wikipedia is the most popular site on the Internet—according to the sampled traffic data. (Source: Wikipedia/Alexa.)

The different metrics for growth all tell a similar story. Until 2002, Wikipedia evolution was driven by a relative handful of insiders and enthusiasts. The mechanism for growth was erratic, and the relative variation in any performance measure, compared to the average level, was fairly high. Sometime after 2002, it appears to have reached a critical mass that drove something similar to self-sustaining exponential growth with respect to almost any metric one wants to consider. Then, towards the end of 2006, the growth fell to a subexponential rate, possibly reflecting saturation of the pool of volunteer contributors, or completion of topic areas that enthusiasts wanted to pursue, or diminishing novelty, or the inevitable loss of perceived prestige ("coolness") when the number of contributors is very large.

The first phase of growth is only hinted at in the figures, since for many metrics the time scale does not extend before 2002. But from a management standpoint, it seems inevitable, and the next section will discuss the role of management in more detail. The exponential growth phase is well supported in all of the figures shown. The third phase is very recent, and the long-term trend cannot be discerned from the available information. It is certainly possible that growth will tick back up, especially if Wikipedia leadership introduces new functionalities that attract fresh

| Rank | Site | Reach | Views | R/W | V/W |
|---|---|---|---|---|---|
| 1 | yahoo.com | 26.8 | 6.0 | 4.25 | 13.3 |
| 2 | msn.com | 30.1 | 3.6 | 4.78 | 8.0 |
| 3 | google.com | 25.1 | 2.3 | 3.98 | 5.1 |
| 4 | youtube.com | 8.9 | 1.6 | 1.41 | 3.6 |
| 5 | myspace.com | 4.4 | 2.4 | 0.70 | 5.3 |
| 6 | live.com | 14.6 | 0.63 | 2.32 | 1.3 |
| 7 | baidu.com | 6.1 | 1.2 | 0.97 | 2.7 |
| 8 | qq.com | 5.3 | 0.73 | 0.84 | 1.6 |
| 9 | orkut.com | 2.6 | 1.4 | 0.41 | 3.1 |
| 10 | wikipedia.com | 6.3 | 0.45 | 1 | 1 |
| 11 | yahoo.co.jp | 2.8 | 0.93 | 0.44 | 2.1 |

**Table 1.1.** Snapshot of the top traffic sites, as estimated for the week of March 2, 2007. The columns "Reach" and "Views" (pages viewed) are percentages of the estimated totals. Columns R/W and V/W are relative to Wikipedia's corresponding statistics. (Source: Wikipedia/Alexa.)

contributions. But it seems more likely that the rapid early spurt is in the past, and the new management model should be one of consolidation and steady, but not explosive, growth.

### 1.3.1 Micro-Growth

Besides the long-term growth phases, there is also interesting variation that occurs on short time scales. There are clear holiday effects in the submission of contributions, and a regular drop in September that people speculate is associated with the distractions attending the start of the school year (cf. `http://en.wikip⊕edia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`).

Also, there are claims that the revision times for Wikipedia constitute a self-similar process (Almeida, Mozafari, and Cho, 2007). Such processes exist in Internet traffic, but the mechanism for such behavior in Wikipedia postings, though provocative, is unclear.

| Statistic | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|
| No. of articles | 11k | 90k | 168k | 379k | 791k | 1.4M |
| No. of internal links | 87k | 1.1M | 2.7M | 7.0M | 16.7M | 32.1M |
| No. of external links | 2.7k | 20k | 76k | 300k | 996k | 2.6M |
| No. of words | 2.4M | 26.2M | 52.0M | 121M | 289M | 609M |
| No. of images |  | 3.9k | 24k | 122k | 388k | 876k |
| No. of contributors | 238 | 1077 | 4282 | 17542 | 56142 | 151934 |
| Contributors (> 5 edits) | 110 | 324 | 1122 | 4853 | 14923 | 43001 |
| Contributors (> 100 edits) | 10 | 100 | 198 | 779 | 1964 | 4330 |
| Mean edits per article | 1.8 | 4.8 | 9.2 | 15.7 | 24.1 | 38.0 |
| No. of categories |  |  |  | 23k | 76k | 176k |
| Categorized articles |  |  |  | 61% | 80% | 86% |
| No. of Wikipedias | 1 | 7 | 17 | 48 | 76 | 113 |

**Table 1.2.** The quoted statistics were measured at the end of October in each of the indicated years. The sources are `http://stats.wikimedia.org/EN/Tables` ⊕ `WikipediaEN.htm` and `http://stats.wikimedia.org/EN/TablesArticlesTot` ⊕ `al.htm`.

## 1.4 CREATING AND CORRECTING CONTENT

From the standpoint of e-commerce, managers want to understand how to replicate the growth mechanisms of Wikipedia. There is no explicit recipe for exporting its success, but some general principles are evident. This section also considers strategies for ensuring future growth through the creation of new kinds of value, as has happened at various times during the evolution of the Wikipedia.

Below the take-off point, we believe that management had to invest significant resources in creating infrastructure, content, and enthusiasm. This section looks at some of those topics in detail. The other main feature of Wikipedia is the flatness of the organization and relatively permissive power-sharing. This fostered a sense of community among the contributors that stands apart from traditional corporate environments and helped to engage and empower a broad base of user-contributors.

### 1.4.1 The Number of Contributors

Over time, Wikipedia's increasing popularity drew more and more users to contribute, creating a positive feedback loop on the quantity and quality of its content. Table 1.2 hinted at some of this. On the one hand, the number of active contributors (users who contributed more than five edits) as a fraction of the total number contributors (users who contributed at least one edit) stabilizes over time. This, together with the observation that the pool of contributors is growing "exponentially" fast suggests that modeling the reach of Wikipedia as a multiplicative process, with saturation, is reasonable. On the other hand, Table 1.2 shows that the pool of contributors who are *very* active is much smaller. The number of very active contributors (users who contribute more than 100 edits), as a fraction of the total number of contributors, has been slowly decreasing since 2003. This suggests that there was an early stage of forced growth driven by management investment in and cultivation of highly active contributors, but that Wikipedia has now reached a point such that the initial pool of these supererogatory contributers is exhausted.

To explore this further, Fig. 1.5 shows that a very small number of people make a great many contributions, but that a lot of people make a small number of contributions. The very smooth curve strongly suggests that a simple behavioral law governs this relationship. If so, then business plans for creating collaborative content can anticipate specific distributions for the degree of volunteer involvement (the parameters of the curve may depend upon the project, but if there is a generalizable behavioral law, the properties of the distribution should be stable). It is likely that the super-contributors are wordsmithing to enforce style conventions, and in a commercial enterprise they would require special compensation or recognition.

### 1.4.2 The Overhead for Content Maintenance

As of January 1, 2007, Wikipedia had 2,463,839 pages related to administration and 3,806,878 pages related to content. The latter includes 855,427 pages of content discussion, in which contributors point out gaps or raise questions about accuracy or style. Clearly, there is a significant overhead associated with the generation and maintenance of Wikipedia content.

For example, there exist nine different types of pages that have the title 'Censorship', including a content page, a content discussion page, a category page, a

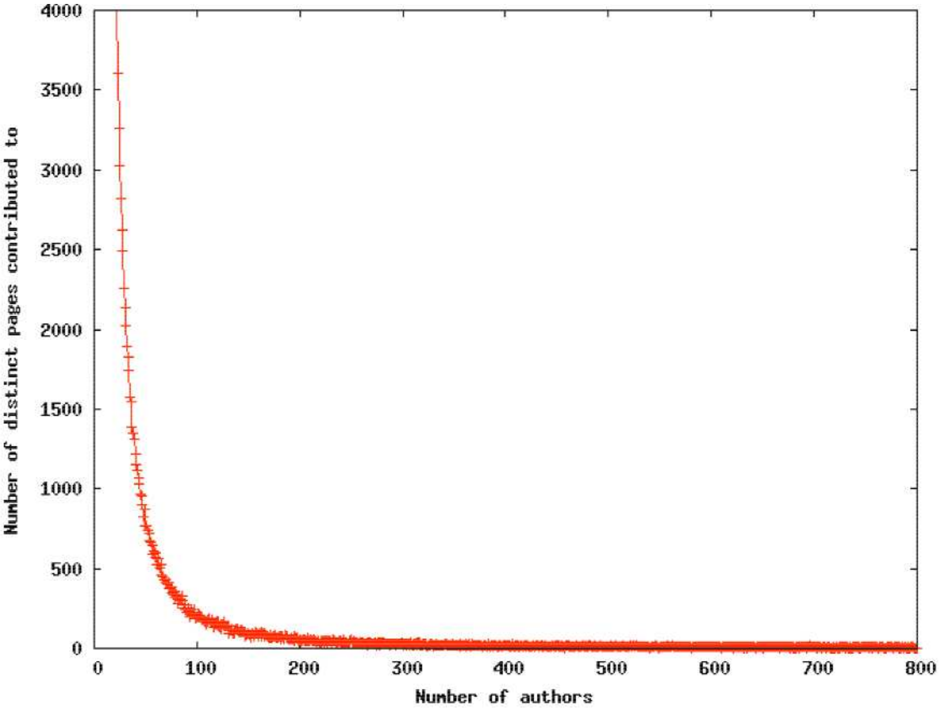

**Figure 1.5.** Histogram of the number of articles contributed to by individuals.

category discussion page, a template page, a template discussion page, a (deleted) user's discussion page, a demonstration page, and the discussion about the demonstration page. All of these pages deal with different facets of the content and its maintenance. This is an extreme example; most topics have fewer pages. But it illustrates the complexity of the underlying structure.

Importantly, there is much less revision activity on the administrative pages than on the content pages. For the content pages, there were 58,636,873 revisions; for the content discussion, there were 6,049,233 separate postings/revisions. In contrast, the administrative pages had only 15,715,896 revisions. In both cases most revisions are relatively minor, but it certainly appears that the administrative pages are more stable and persistent than the content pages. The discussion pages are more complex; some topics are controversial and generate a great deal of discussion, but most get little attention. But the short message is that 61% of the pages carry content, but 80% of the revisions are about content. From a business standpoint this seems like the administrative pages carry significant overhead, but note that Wikipedia has successfully offloaded much of that burden to a decentralized volunteer community.

To assess the balance between content creation and content maintenance, we examine the relationship between the number of discussion points on a topic and the number of content revisions. We found that a unit of discussion produces slightly less than a unit of content, both at the aggregate level for all of Wikipedia and for the topic category of mathematics. Figure 1.6 plots the amount of content and the amount of discussion for a random sample of 50,000 topics. The 45-degree line corresponds to the case in which one discussion entry corresponds to one content revision.

The vertical lines in Fig. 1.6 mostly correspond to holding pages in high-level categories where new articles are entered prior to indexing and linking. For these pages the content changes rapidly, but there is relatively little discussion. Note that this kind of enrollment process requires regular management and attention from domain experts.

As a comparison, we also created a similar plot for the first 200 Wikipedia articles indexed under mathematics. We chose this topic because it was numerous and because mathematics is not intrinsically controversial, in the way that articles on politics, history, or *Star Trek* might be. Surprisingly, the same general pattern noted for the sample of 50,000 articles from all of Wikipedia holds for the 200 articles in mathematics.

These findings are compatible with two quite different views of the database. Either the high quality of first drafts leaves little room for disagreement, or the amount of content supervision is very low, or both.

### 1.4.3 Content Protection

In any business plan with content creation by an open community, it is an explicit management responsibility to protect content quality. In Wikipedia, the most visible aspect of this problem is protection of content from vandalism. Section 2 listed some of the more famous instances of deliberate content destruction. Wikipedia has two main defenses: they can restore the original content, or they can freeze the entry, so that unauthorized contributors cannot change the contents.

To get a sense of the scope of the problem, Fig. 1.8 plots the number of times that inappropriate content has been deleted from the Wikipedia system (plotted by

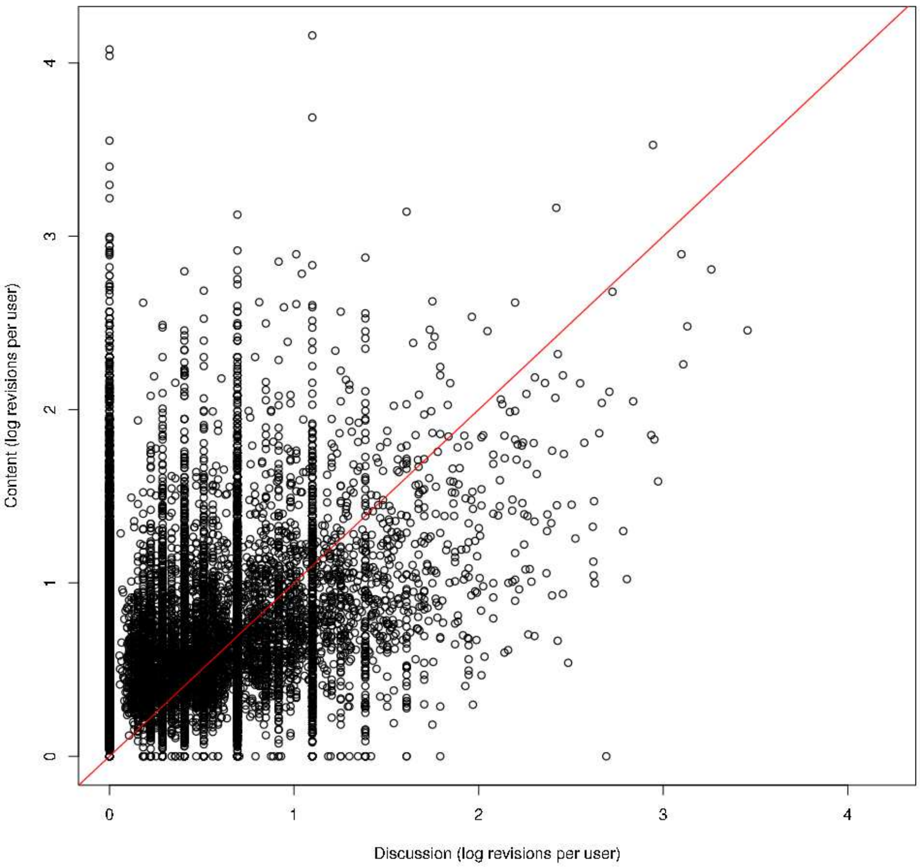

**Figure 1.6.** This shows the amount of revision of discussion pages compared with the number of revisions of content pages for 50,000 topics. (The 45-degree line is shown for convenience in comparison.)

×), as well the number of times that contents were protected from future changes (plotted by +). Both are good indicators of the amount of vandalism present in the system. The increase over time in these variables is linear, rather than the exponential trend seen in almost every other plot of Wikipedia activity. This probably reflects the fact that this deletion and protection are administrative tasks that require executive attention, which is a limiting resource. And it is worth noting that many minor instances of vandalism are probably never noticed.

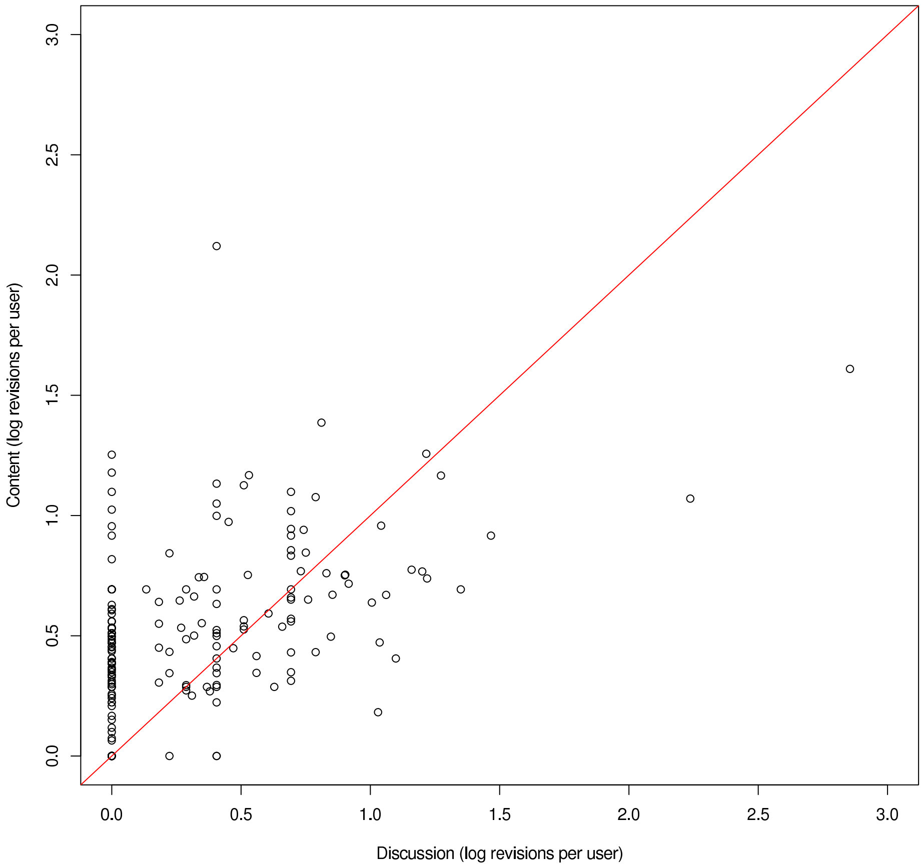

**Figure 1.7.** This shows the amount of revision of discussion pages for 200 mathematics articles compared with the number of revisions of the corresponding content pages. (The 45-degree line is shown for convenience.) As before, a unit of discussion produces slightly less than a unit of content.

For emerging businesses, it would make sense to develop a text-mining system that scrutinizes entries or edits for possible violations of the social compact. Certainly there are keywords that help flag problems, and revisions by anonymous contributors could also raise cautions. More subtle signals are also possible; the success of Bayesian methods for spam filtering (Madigan, 2005) suggests that a great deal of progress is possible.

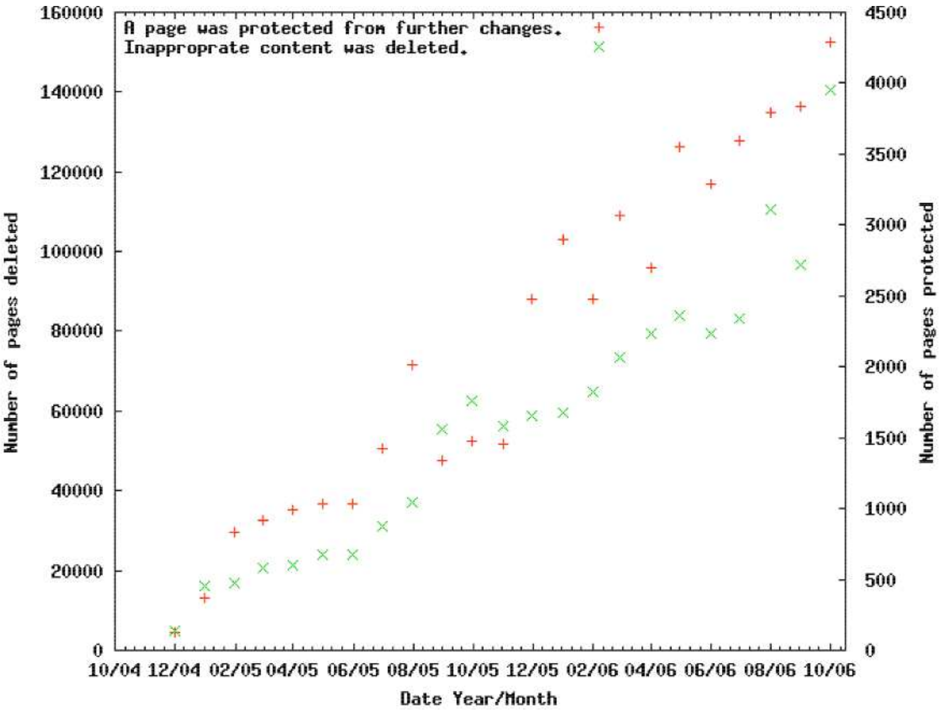

**Figure 1.8.** Amount of content deleted and protected over time. The '+' indicates the number of protected pages, and the '×' indicates the number of times content was deleted.

A related topic concerns effective management strategies for content control. We do not know what criteria Wikipedia administrators use in deciding whether to freeze a particular entry, but it is likely that whatever (possibly informal) guidelines exist depend upon the history of attacks upon the text. Figure 1.9 shows a frequency plot of the number of times the content of a page was protected and the number of times the page was revised. It appears that most protections occur when there has been no revision; these are probably administrative pages, or obviously controversial topics. A small number of protections occur after substantial revisions; these may represent honest intellectual disagreements for which a community consensus slowly crystalizes. In between is the domain in which management policies might have impact; for example, a policy of protecting any entry that is vandalized once would be a reasonable approach.

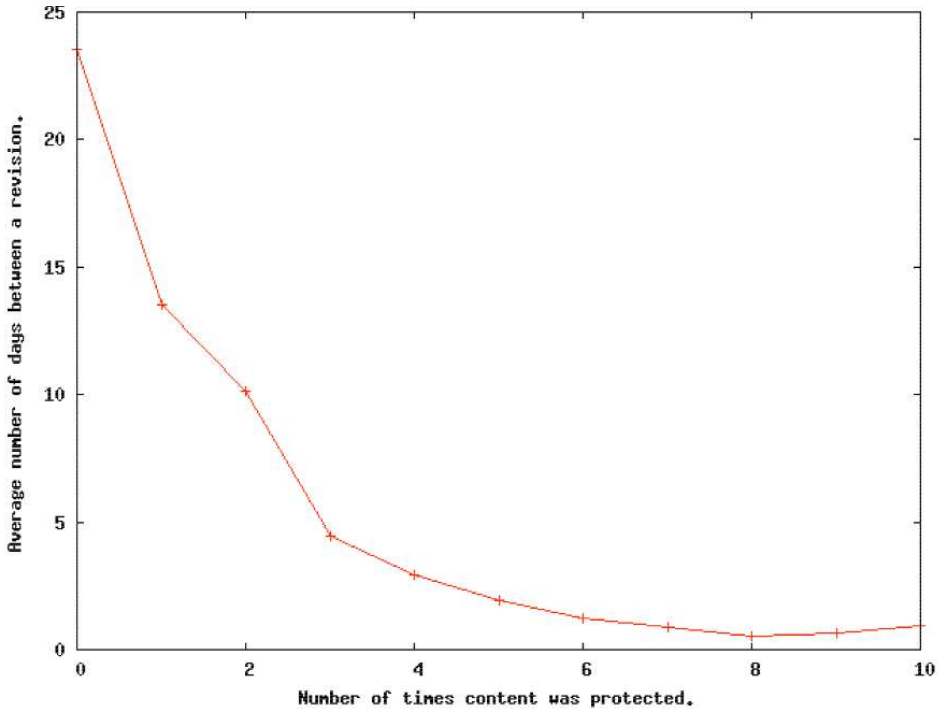

**Figure 1.9.** Relationship between the rate of revisions and the number of times that content was protected.

### 1.4.4 Revision Management

Revision is a balancing act when handling user-contributed content. If there is too much, then the product is unstable. People want to know that what they saw a month ago is probably still there. But if there is too little revision, then some of the key benefits of open collaboration are lost. Wikipedia has tended to indulge revision, and this has worked well, but how well this policy generalizes to other applications is unclear.

In order to study the role of revision, we considered a subset of 900,000 randomly chosen articles. For each page we counted the number of revisions to the discussion page, the number of unique users participating in the discussion, the number of revisions to the content page, and the number of unique users providing content. We found that, on average, a modification to to a page occurs every 23 days with

7 different individuals providing inputs over 13 revisions. But the tail behavior is extreme.

In Fig. 1.10 each point corresponds to an article. The $x$-axis measures the number of revisions to the content, and the $y$-axis measures the number of unique contributors to the discussion and/or the revision. As it shows, it is not uncommon for some articles to be touched by hundreds of hands. At the same time, the number of revisions tends to be more than the number of participating users. This points up the wisdom of the *laissez-faire* approach taken by Wikipedia; people make many small edits, and in general the quality improves.

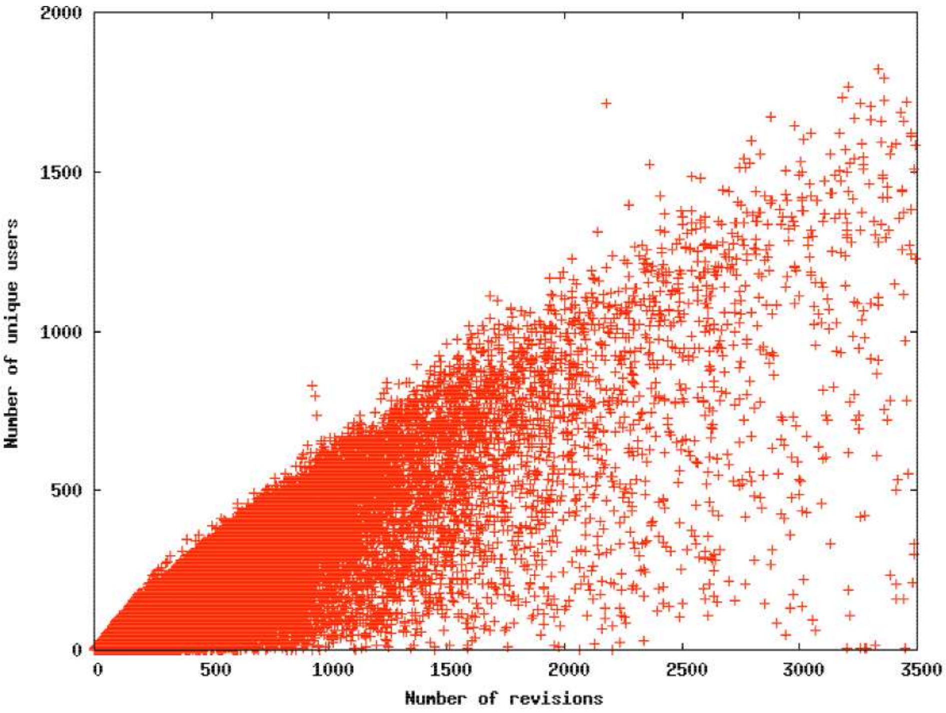

**Figure 1.10.** This shows the number of contributors as a function of the number of revisions. Note that some entries have thousands of revisions and contributors; it is likely that many of these are pages with lists.

Obviously, for most articles the rate at which changes are made should diminish over time as errors are removed, key facts are validated, consensus is reached, and the attention of the administrative community shifts to other topics. Figure 1.11 shows how the number of actively edited articles and stale articles changes over time (we define an active article as one that has been revised within the last 6 months). We also plot the number of active contributors over time.

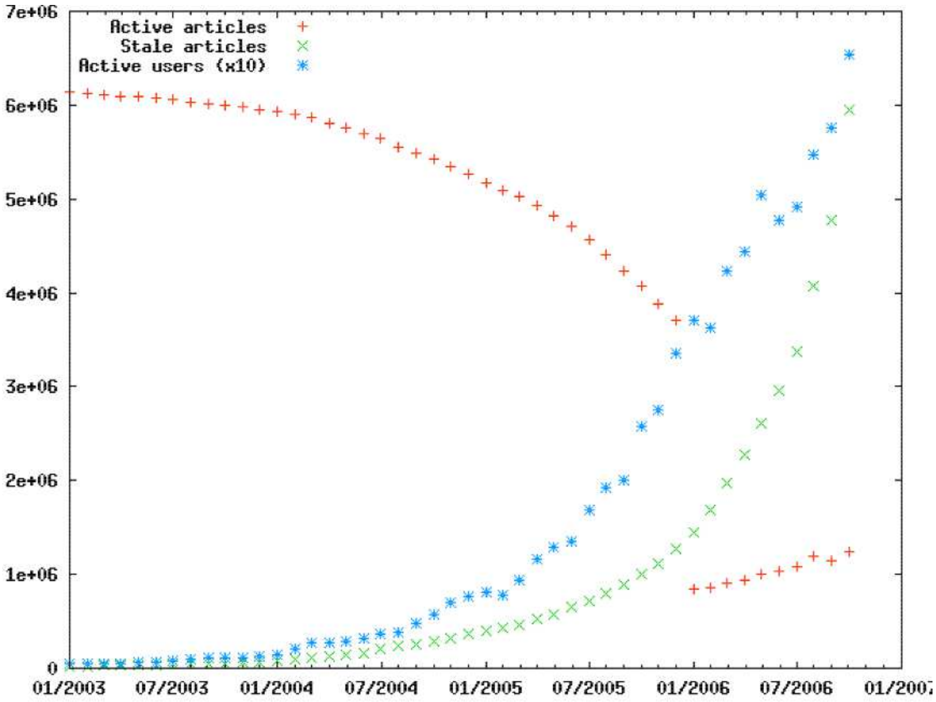

**Figure 1.11.** Tracing the number of entries being actively edited (+), the number of entries no-longer being edited (×) and the number of active contributors (*) over time.

The behavior of the number of stale entries and the number of active contributors is reasonable. But we have no explanation for the sudden drop in the number of actively edited entries around January 1, 2006. We suspect this indicates some kind of change in the management of the revision process, and we hope that Wikipedia insiders can clarify this.

### 1.4.5 Linkages

A key functionality that Wikipedia provides is internal links between articles and links to external webpages. Obviously, the latter pose a management problem in terms of potential instability, but the convenience has more than compensated for the occasional dead link.

To get a sense of the scale, as of January 1, 2007 there were 2,528,868 articles with outside links, 87,593,800 non-duplicative links between articles within Wikipedia, and 14,079,567 category links that were explicitly defined. The number of category links that were also represented as content-to-content links was 813,176. Thus, there has been significant effort by contributors to differentiate between the semantic linkages of the content and the taxonomy used to classify it. Obviously, this has implications for a business model that includes multiple kinds of links.

In terms of designing such systems, there is the question of whether the taxonomy should be provided mostly by administrators or mostly by the content providers themselves in a self-organizing manner. For Wikipedia, traversing the category linkages from the top-level concepts shows that only about 10,000 categories are related to administrative matters, whereas the remaining categories were generated by the end-users themselves. Thus it would seem that within the Wikipedia experiment the taxonomy being used is not only created on-the-fly by the community, but it is also constructed by a community that clearly differentiates between content and classification.

For Wikipedia, many articles now include user-contributed links to other websites on the Internet. Specifically, 535,750 content pages link to one or more URLs outside the Wikipedia namespace. This means that about 81% of Wikipedia content pages have no links to the outside Web. However, it could still be the case that most of the content pages are actually pointers to other pages, with little actual written prose to analyze. In this case, 1,538,983 (or slightly more than half of the content pages) reference at least one other Wikipedia page. Hence the data within Wikipedia seems to be highly self-referenced, while still containing substantial user-generated content.

For additional perspective on the linkage patterns within Wikipedia, Fig. 1.12 plots the relationship between the length in characters of a Wikipedia page and the number of web links to pages outside of Wikipedia. The figure suggests a superposition of several different clusters of pages. The pages that have small numbers of characters but many internal links are probably administrative lists; pages that are less extreme, but still have many links per character, could be indexes. The structures suggested in the plot require further study, but indicate a wide range of internal-link behavior.

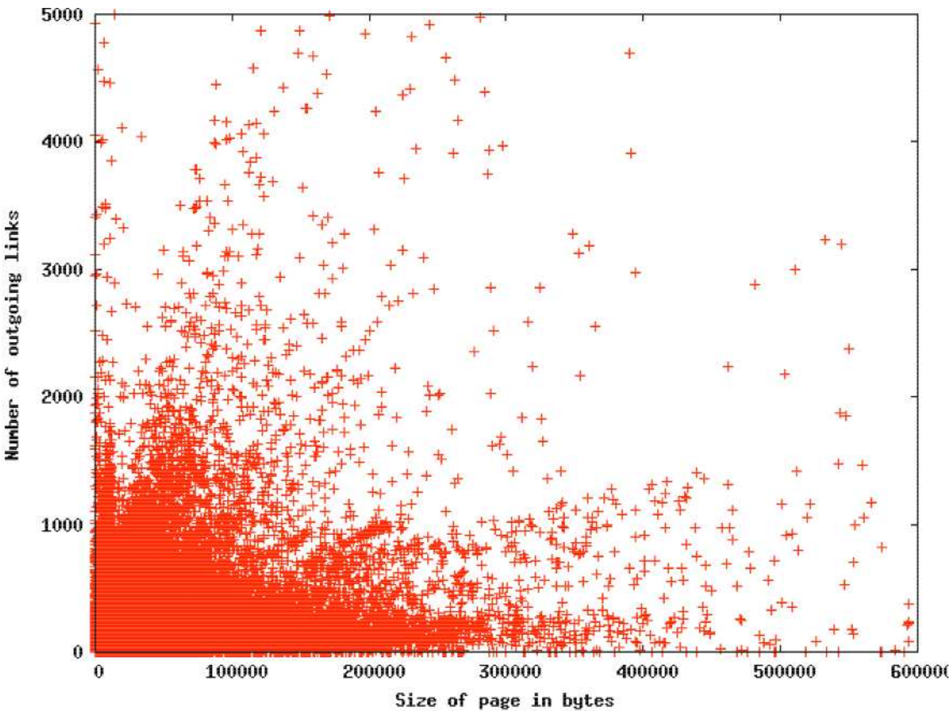

**Figure 1.12.** Plot of the relationship between the number of outgoing links to other pages and page length.

Symmetrically, Fig. 1.13 plots the relationship between the number of external links to the Internet on a page and its length in characters. Note that this figure shows even more structure, with clearly defined rays emanating from the origin. The most vertical ray probably corresponds to sites that are essentially lists of URLs, but the other rays invite further study. It certainly appears that there are distinct clusters of pages with respect to external links.

Note that both of these figures contain artifacts that imply a linear relationship between the number of links and the size of the page. The reason for this is that the links themselves occupy space within the page and therefore the number of links directly influences the size of the page. This is more evident in Fig. 1.13 because the average length of a URL is 58 characters, versus an internal Wikipedia link that is usually 18 characters long. Nonetheless, these artifacts only partially account for the observed structure.

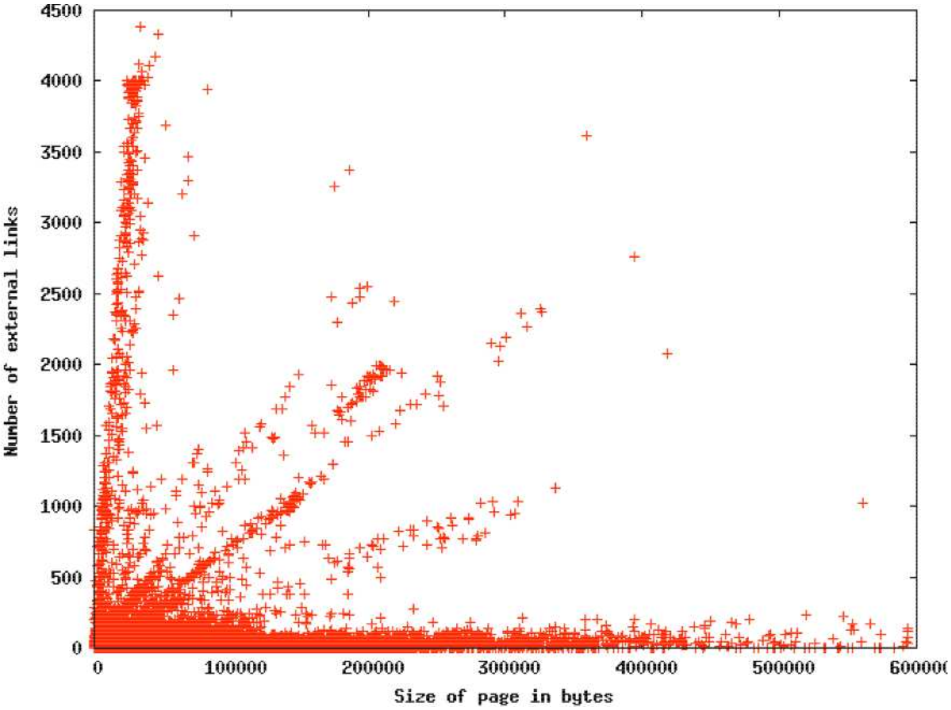

**Figure 1.13.** Plot of the relationship between the number of external links to Internet and page length.

To complete the narrative loop, Fig. 1.14 plots the number of external links against the number of internal links. There are many pages that have large numbers of external URL links but few internal links. As a generalization, pages with very large numbers (over 1,000) of external links tend to be automatically generated tools-pages meant to be used for maintenance. Pages with significant numbers of links (100s) tend to be lists of places, people or objects. Again, there is clear structure in the graph, but only partial explanation.

To close, consider the external perspective. Figure 1.15 shows the relationship between the number of external pages that point to a specific Wikipedia article and the number of links from that article to pages outside the Wikipedia namespace. The mass at the left is simple to understand; many Wikipedia pages have lists of external links but do not have pointers from the outside. The long-tail to the right is surprising.

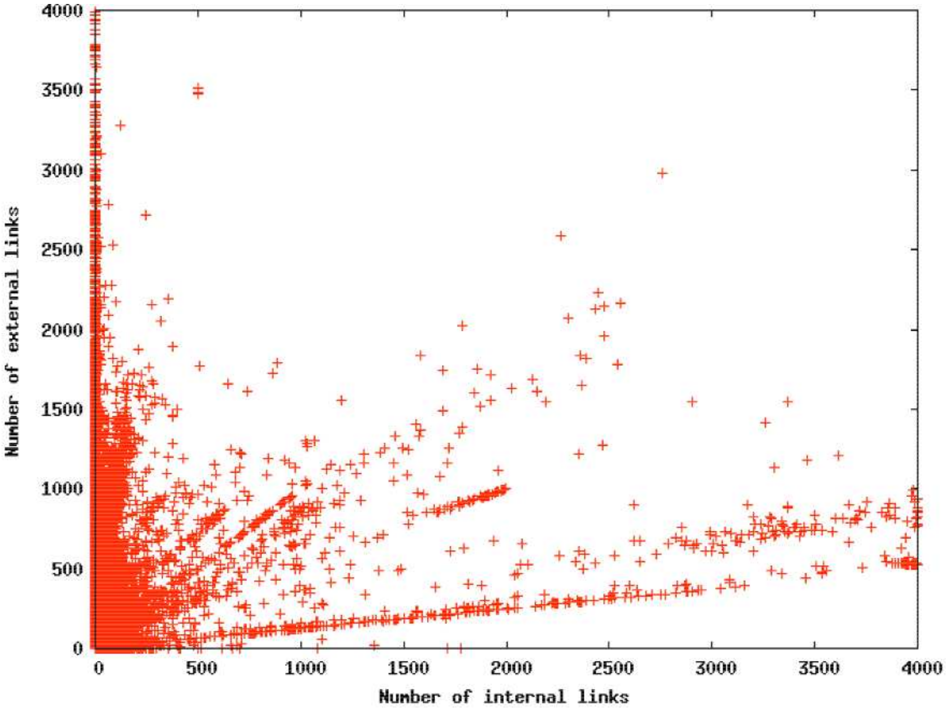

**Figure 1.14.** Plot of the number of links external to Wikipedia against the number of internal links for a sample of Wikipedia articles.

### 1.4.6 New Functionality

As with any successful enterprise, the Wikipedia has not only grown—it has changed. Key innovations were the enabling of categories, external links and the extensive use of images,all of which emerged after the initial format had been developed. It seems likely that audio and video capability will someday be added, as well as transparent links to external software packages.

But other kinds of functionality may have even larger implications. Alexander Wissner-Gross, a Ph.D. physics student at Harvard, has developed software to help Wikipedia users find related information on a general topic. The algorithm uses text mining, as well as information on the popularity of particular paths through the Wikipedia network of links. In some sense, this is rather like the popular recommender system used by Amazon to point out books a customer might enjoy, based on their purchase and browsing history.

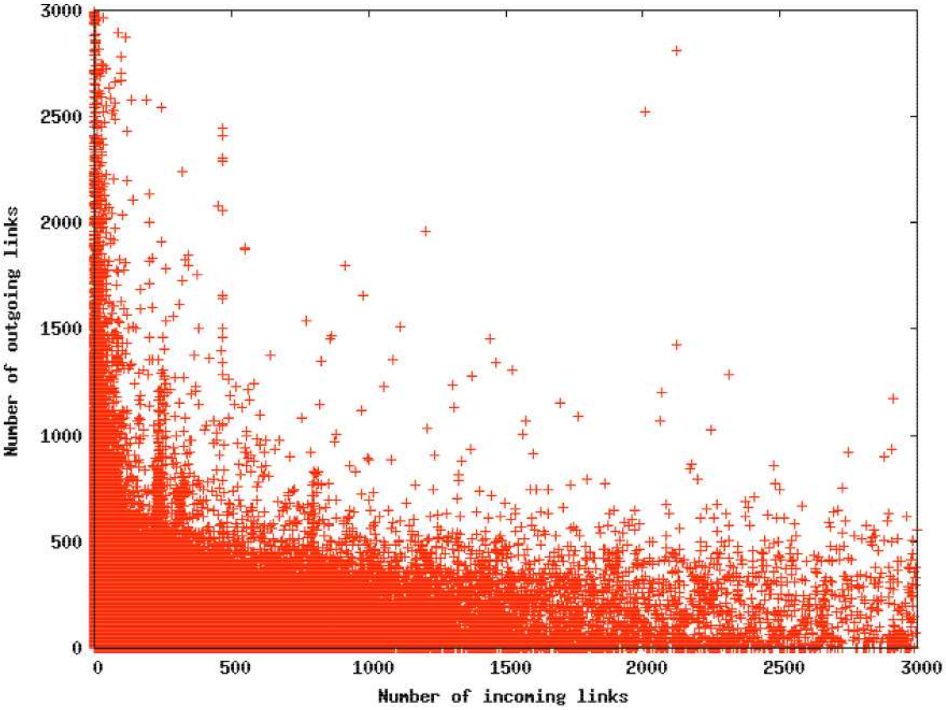

**Figure 1.15.** Comparison plot of the relationship between the number of incoming and outgoing links to other pages within Wikipedia.

Also, Luca de Alfaro at the University of California at Santa Cruz has developed software that estimates, phrase by phrase, the trustworthiness of text in Wikipedia articles (Powell, 2007). The procedure is based on tracking the number of times that a particular content contributor's work has been removed or revised. Color coding flags text by people with high rates of reversal as potentially less reliable than other portions of the same article. This capability is another example of the fresh and unforeseen potential of the Wikipedia data archives.

More broadly, the organic growth of links between Wikipedia topics demands network analysis. There are probably deep questions about information structures that could be addressed. For example, what are the empty spots in the Wikipedia system, and how would one notice them? Do different fields have similar internal connection structures, or do some fields show very different kinds of linkage? How might one segment the Wikipedia network into meaningful cliques, and would these correspond to a recognizable ontology?

With regard to clique segmentation, there are a number of traditional approaches developed in the social network community, but these are mostly ad hoc. If one wanted to estimate the size of the clique corresponding to, say, mathematics, it would useful to adapt methods developed for estimating the size of the World Wide Web (Dobra and Fienberg, 2003; Bradlow and Schmittlein, 2000) that are based on capture-recapture models and Markov chain explorations.

## 1.5 DISCUSSION

Although the Wikipedia is not a for-profit enterprise, it is a unique example of a novel approach to constructing value. As such, its evolution and management structures hold important lessons for e-commerce.

In the first part of this paper, we focused on the growth history of Wikipedia. The mathematical picture of exponential growth in the middle phase is well-established, according to many different metrics of growth. In the late phase, there is emerging evidence that growth has become subexponential, and the causes for this (aside from mathematical inevitability) are unclear. Bold development of new functionality could easily re-establish exponential growth for a while. Our data have least to say about the first phase, during which the Wikipedia founders established the infrastructure and recruited an initial team of enthusiastic content creators. But it seems clear that critical ingredients were a social network within the encyclopedia community, building on the Nupedia connections, and a flat, decentralized management system that invited self-paced contribution and recognized volunteerism.

The second part of the paper focused on the technical mechanisms of content creation. This addressed growth in the number of contributors, the administrative costs of content maintenance (as inferred from administrative pages), the balance between open editing and content protection (as indicated by trends in the number of protected pages and editing histories), revision management, the different kinds of links needed to support the Wikipedia functionalities, and prospects for new kinds of service in the future.

As a research area, Wikipedia science is exciting. There is an enormous amount of data, and whenever one looks closely there are research problems. It is rich example of the evolution of a self-organizing system, and its processes inform many aspects of organizational theory.

### Acknowledgement

Robert Warren is partially supported through NSERC and Ontario Graduate Student Scholarships. Edoardo Airoldi is supported by NIH grant R01 GM071966 and NSF grant IIS-0513552, both to Olga Troyanskaya at Princeton University. David Banks was partially supported by NSF grant DMS-0437183.

## 1.6 REFERENCES

Almeida, R., Mozafari, B., Cho, J. (2007). "On the Evolution of Wikipedia," International Conference on Weblogs and Social Media, `http://www.ic` ⊕ `wsm.org/papers/paper2.html`.

Bradlow, E., and Schmittlein, D. (2000). "The Little Engines that Could: Modeling the Performance of World Wide Web Search Engines," *Marketing Science*, **19**, 43–62.

Dobra, A., and Fienberg, S. (2003). "How Large Is the World Wide Web?" In *Web Dynamics*, M. Levene and A. Poulovassilis, editors, Springer-Verlag:New York, 2003, pp. 23–44.

Giles, J. (2005). "Internet Encyclopedias Go Head to Head," *Nature*, **438**, 900–901.

Holloway, T., Božičević, M., and Börner, K. (2006). "Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors," `http://arxiv.o` ⊕ `rg/pdf.cs.IR/0512085`.

Leuf, B., and Cunningham, W. (2001). *The Wiki Way: Quick Collaboration on the Web*, Addison-Wesley, New York NY.

Leyden, J. (2006). "Wikipedia Blaster 'Fix' Points to Malware," 11/3, `htt` ⊕ `p://www.theregister.co.uk/2006/11/03/wikipedia_blaster_attack/`.

Madigan, D. (2005). "Statistics and the War on Spam," in *Statistics, A Guide to the Unknown*, ed. R. Peck, G. Casella, G. Cobb, R. Hoerl, and D. Nolan, Duxbury-Brooks/Cole:Belmont CA, pp. 135-147.

Pava, A. (2006). "Colbert Banned from Wikipedia," Civic Actions, August 2, `http://www.civicactions.com/node/405`.

Powell, H. (2007). “New Program Color-Codes Text in Wikipedia Entries to Indicate Trustworthiness,” U.C. Santa Cruz press release, `http://www.u` ⊕ `csc.edu/news_events/press_releases/text.asp?pid=1471`.

Reuters (2006). “Intelligence Czar Unveils Spy Version of Wikipedia,” `htt` ⊕ `p://news.zdnet.com/2100-1009_22-6131309.html`.