Redundancy and Collaboration in Wikibooks

Ilaria Liccardi[2,1], Olivier Chapuis[1,2], Ching-Man Au Yeung[3], and Wendy Mackay[2,1]

[1] Univ. Paris-Sud & CNRS, Orsay, France
[2] INRIA, Orsay, France
[3] NTT Communication Science Laboratories, Kyoto, Japan

Abstract. This paper investigates how *Wikibooks* authors collaborate to create high-quality books. We combined Information Retrieval and statistical techniques to examine the complete multi-year lifecycle of over 50 high-quality Wikibooks. We found that: 1. The presence of redundant material is negatively correlated with collaboration mechanisms; 2. For most books, over 50% of the content is written by a small core of authors; and 3. Use of collaborative tools (predicted pages and talk pages) is significantly correlated with patterns of redundancy. *Non-redundant* books are well-planned from the beginning and require fewer talk pages to reach high-quality status. *Initially redundant* books begin with high redundancy, which drops as soon as authors use coordination tools to restructure the content. *Suddenly redundant* books display sudden bursts of redundancy that must be resolved, requiring significantly more discussion to reach high-quality status. These findings suggest that providing core authors with effective tools for visualizing and removing redundant material may increase writing speed and improve the book's ultimate quality.

Keywords: Collaborative writing, text redundancy, coordination mechanisms.

1 Introduction

The advent of the World Wide Web and wiki-based collaboration technologies has made it possible for a new form of mass collaboration in which groups of strangers work together on a common topic. These on-line, volunteer-based projects have produced major new resources. One of the most successful examples is the Linux kernel, which was developed by a large number of unpaid contributors [27]. More recently, wiki technologies have been introduced to facilitate the creation and editing of interlinked pages, e.g. the Wikipedia encyclopedia and Wikibooks. Unlike collaborative writing within corporate environments, wiki technologies permit large numbers of strangers from around the world to work together on a shared topic.

This type of collaborative writing has inherent advantages and disadvantages. Each project may benefit from a wealth of expertise and knowledge but faces enormous coordination and communication challenges in order to produce a coherent final result. Manuscripts typically evolve during the writing process and discussions and disagreements inevitably occur, due to differences in knowledge, experience and points of view. Groups of co-authors manage their work differently, which affects writing speed, manuscript structure, the validity of the arguments and the perceived quality of the text itself.

In traditional collaborations, contributors are linked by professional ties. Thagard [31] identifies three types of collaboration that reflect the backgrounds and roles of the authors: *dominant relationships*, e.g. employer/employee or teacher/apprentice, *peer relationships*, e.g., among researchers with similar backgrounds, knowledge, and skills, and *peer-different*, e.g., among researchers from different disciplines who share similar goals. However, even in corporate environments, studies of informal collaboration [23, 24] show that motivated individuals sometimes voluntarily take on coordination tasks to support their colleagues.

Unlike projects that occur within a corporate hierarchy, open-source collaborations are often driven by what Lerner calls ”hobbyists” [19] who do not have clearly defined roles with respect to each other. Thagard’s classifications are particularly difficult to define when authors do not interact directly with each other. This raises the question of how large-scale, volunteer-driven writing projects can coordinate the efforts of large numbers of users and still produce high-quality results.

Researchers in CSCW have begun studying large-scale writing projects, hoping to gain insights into how best to design tools to facilitate collaboration. To date, Wikipedia is most well-studied [14, 16, 22]. Researchers have examined communication patterns, conflict resolution and authorship and has used some of these findings to design tools to support collaboration [3]. They disagree about the benefits of including very large numbers of participants. On the one hand, new authors offer the potential for gaining additional expertise and novel perspectives, thus increasing the value of the result [4, 11]. However, adding authors may also reach a point of diminishing returns, with a trade-off between the benefits of additional resources and the costs of increased coordination [10].

Brook’s Law [1] famously argues that additional coordination costs can easily overwhelm any benefits from added personnel: “Adding manpower to a late software project makes it later”. For this reason, corporate managers and book editors often choose to limit group size to increase productivity [29]. In stark contrast is Raymond et al.’s [27] claim that involving as many authors as possible improves open-source software projects. They suggest that it is important to “delegate everything you can, be open to the point of promiscuity”. Kittur and Kraut [14] refine this claim, showing that coordination is essential for harnessing the wisdom of the crowds. They studied the relationship between the number of authors and the use of appropriate coordination tools on the quality of Wikipedia content. Essentially, articles with many authors are better, but only if the authors can effectively coordinate their activities. Of course, this also depends upon the type of work. Stewart [30] shows that larger teams generally perform better when they engage in low-coordination as opposed to high-coordination work.

We are interested in a less well-known, but no less interesting, collaborative writing system, called Wikibooks. The primary goal of Wikibooks is to provide free, printable textbooks that can be used in the classroom. Although most are educational [28, 34], they also cover a wide range of other topics, including sports, religion, interpersonal relationships and even a guide to Harry Potter novels.

Wikibooks is based on the same MediaWiki software and open editing policy as Wikipedia. However, Wikibooks co-authors face a greater challenge than Wikipedia contributors, because they must coordinate their activities on a much greater scale, over much longer periods of time. The longest Wikipedia article is around 50,000 words whereas over 100 Wikibooks exceed this.

In order to study how large groups of strangers coordinate their activities over long periods of time, we examined the first complete set of Wikibooks logs, dating from its creation in 2003 until 2009. Our goal was to identify the key coordination and communication mechanisms that affect quality.

Our first challenge was to find appropriate measures of quality. Studies of Wikipedia have identified a diverse set of possible metrics, including: number of edits and unique editors [22], factual accuracy [8] (but disputed by [6]), credibility [2], revert times [33], and the formality of language [5]. In each case, researchers applied these metrics to small samples of Wikipedia articles and, sometimes, to equivalent articles in traditional encyclopedias as an independent standard of quality.

However Wikibooks involve even larger scale collaborative efforts, which renders some of these measures infeasible. We decided to focus on the 59 'featured' Wikibooks, which had been designated of the highest quality, and use both Information Retrieval and statistical techniques to analyze the historical data. We focused on four key measures: level of redundancy in the text, authorship patterns, use of *predicted pages* and use of *talk pages*.

This paper begins with a description of the Wikibooks corpus, followed by the study design and our analysis methods. We then describe and discuss the results of three successive analyses: 1. Redundancy, 2. Co-authorship and 3. Collaboration lifecycle. We then provide an in-depth look at a few specific books and conclude with implications for design and directions for future research.

2 Wikibooks

We analyzed data obtained from the English version of Wikibooks[1]. Since its beginning in 2003, the site has expanded to include over 2000 books, the content of which is contributed entirely by volunteers.

2.1 Corpus

The Wikibooks site includes books in all stages of development. The complete history of each page on Wikibooks is stored in a database on the website and it is possible to access every past revision of a page through the web interface. The Wikimedia Foundation provides downloadable versions of the database including the page history. Our data set comes from a database dump of the Wikibooks Website on 15 May 2009[2] and contains 2,039 books. Because we were interested in understanding the factors that are correlated with high-quality books, we focused our analysis on the 59 Featured books.

Of these 59, we removed eight books: Five were from the Wikijunior series, which are designed to be age-appropriate non-fiction books for children from birth to age 12. These books have an atypical structure, with very little text and many images. The co- authors discuss the content thoroughly before beginning to write to ensure that it is suitable for the intended audience. We also removed FHSST PHYSICS since

[1] http://en.wikibooks.org/

[2] http://download.wikimedia.org/enwikibooks/

this book's content was based on an existing text written outside of the Wikibooks environment. Finally, we removed ADVENTIST YOUTH HONORS ANSWER BOOK and SOCIAL AND CULTURAL FOUNDATIONS OF AMERICAN EDUCATION since they were structured more like catalogs or reference sources rather than actual books. We analyzed several statistics from these books to understand the level of participation in each book from its beginning to the time that the book was designated as a 'featured book'.

2.2 Featured Books as a Measure of Quality

Wikibooks does not provide a quantitative measure for book quality, so we chose an empirical, but qualitative measure, 'featured book' status. The Wikibooks community identifies a small number of books that they judge to be of high quality, based on a set of pre-defined criteria. These criteria are imprecise (probably deliberately so), but do offer guidelines as to what constitutes a good book[3]. Criteria include, for example, the clarity of the text, the structure and completeness of the book, and whether or not the book conforms to Wikibooks policies.

Any Wikibooks reader can nominate a book for 'featured book' status. Once a number of users have so nominated a book, an administrator reviews the strength of the arguments with respect to the above criteria. If it passes this test, the book is voted on through a democratic process. Successful books are then added to the list of 'featured books'[4]. Only a small subset of Wikibooks is nominated and they represent about 3% of the books available on the site. The guidelines are stringent and these books must maintain their quality in order to maintain featured status. Because volunteers may edit books at any time, the Wikibooks community actively monitors the quality of these books and removes them if necessary.

Previous studies of Wikipedia [14, 15, 16] used a similar measure of quality, based on whether an article was explicitly chosen to be featured by Wikipedia administrators. In other studies, Wikipedia users were asked to read an article and rate its quality [13]. However, as mentioned earlier, since Wikibooks are much longer than encyclopedia entries, the latter strategy is not practical for our purposes.

Featured and Non-featured. Because we were interested in understanding the collaborative process involved in writing *successful* books, we focused solely on featured books. We calculated measures of redundancy from the inception of each book through to its completion, as indicated by its election to 'featured book' status. Although these books comprise only a small percentage of the total number books, they are much longer on average and account for almost half of the total size of the Wikibooks database. We recognize that some non-featured books are also close to completion, but since we have no objective, independent criteria by which to measure the quality of such books, we do no include them.

[3] http://en.wikibooks.org/wiki/Wikibooks:Good_books

[4] http://en.wikibooks.org/wiki/Wikibooks:Featured_books

3 Analysis 1: Redundancy Patterns

One possible measure of how authors coordinate their activities is *redundancy*. We hypothesized that redundant text can act as a proxy for communication breakdowns and a lack of coordination among participants. We also expected that redundant text will be highly correlated with other negative indicators, such as the presence of cleanup and maintenance tags. We thus measured redundancy within individual books, looking for patterns of how it changed over time. At regular intervals during the evolution of each book, we took a snapshot and analyzed the full text of the manuscript at that point, and counted the number of redundant paragraphs.

Note that redundancy can only be interpreted within the context of the particular book. Because different Wikibooks are structured differently and cover completely different subject areas, it does not make sense to compare levels of redundancy across books: some books may require some redundancy whereas others do not. Thus, our classification is not based on absolute redundancy values, but rather on an analysis of how redundancy curves change over time, within the context of each individual book.

Research Questions. We are particularly interested in:

- How does the level of redundancy change over time?
- Do different books exhibit different patterns of redundancy from their beginning to completion?

3.1 Measuring Redundancy via Semantic Similarity

We base our measure of redundancy on argument *repetition*, i.e. the amount of similarity among arguments (sentence, recurring words etc.), at the level of paragraphs, sections and chapters, at different points in the writing process. This provides us with an objective, quantitative measure of the effectiveness of the different communication and collaboration mechanisms. Although redundancy is sometimes used for automatic document summarization, in which redundancy in the summary or abstract is used as an indicator of poor quality, we are unaware of other studies that use redundancy as a measure of coordination.

Our strategy is to quantify the level of redundancy by determining the semantic similarity between two sentences. Although this is challenging even with modern natural language processing techniques, a combination of techniques has proven to be effective, e.g. [9, 20], and offers an approximation for the amount of similarity and thus redundancy between two sentences.

Words are the most basic unit of measure for identifying similarity between texts. Techniques for measuring similarity are based on string similarity [12], thesauruses [26] or corpus statistics [32]. In information retrieval, the ‘bag of words’ approach is commonly used to measure similarity between documents. A document is usually characterized by a term vector of length *n*, the elements of which indicate whether a term is present in the document and its relative importance. Terms are usually weighted using the TF-IDF (term frequency-inverse document frequency) scheme [17]. The cosine similarity measure is then used to compare the two vectors.

Other methods have also been proposed that exploit word co-occurrence information instead of using exact word matching. For example, latent semantic analysis [18] can be used to measure similarity between texts by computing higher-order word relations based on dimensionality reduction. Some approaches combine different techniques. For example, Li at al. [21] propose a method that combines Word Net based word similarity, corpus statistics and word order similarity. Islam and Inkpen [12] propose a similar approach based on substring matching of words, point-wise mutual information similarity, and word order similarity.

Redundancy Measure. We chose to use cosine similarity between term vectors constructed by using the TF-IDF weighting scheme [17] to measure redundancy between paragraphs. We decided to use cosine similarity to measure pairwise similarities in each snapshot of each Wikibook.

In formal notation, let T be the set of terms that appear in a Wikibook B. We employ standard Information Retrieval preprocessing methods, such as stop-word removal and stemming [35] to produce the set P of paragraphs. Each paragraph $p \in P$ is characterized by a term vector:

$$v_p = (w_{p,1}, w_{p,2}, \ldots, w_{p,|T|})$$

where $w_{p,i}$ is the weight of term $t_i \in T$ given by the standard TF-IDF weighting scheme.

The similarity between two paragraphs p and q can then be calculated by using the cosine similarity measure, given by:

$$sim(v_p, v_q) = \frac{v_p \cdot v_q}{\|v_p\| \times \|v_q\|}$$

The cosine similarity tells us only how similar two paragraphs are with respect to the terms they contain, but does not tell us how much redundancy is observed in the book. Here, we define redundancy as the proportion of pairs of paragraphs that attain a similarity value higher than a threshold α :

$$redundancy(B) = \frac{\sum_{p,q \in P, p \neq q} \delta(p,q)}{2 \times C_2^{|T|}} \quad where\ \delta(p,q) = \begin{cases} 1 & if\ sim(v_p, v_q) \geq \alpha \\ 0 & otherwise \end{cases}$$

and since there are $C_2^{|T|}$ pairs of (p,q) and the similarity function is symmetric, the term $2 \times C_2^{|T|}$ is used to normalize the redundancy score.

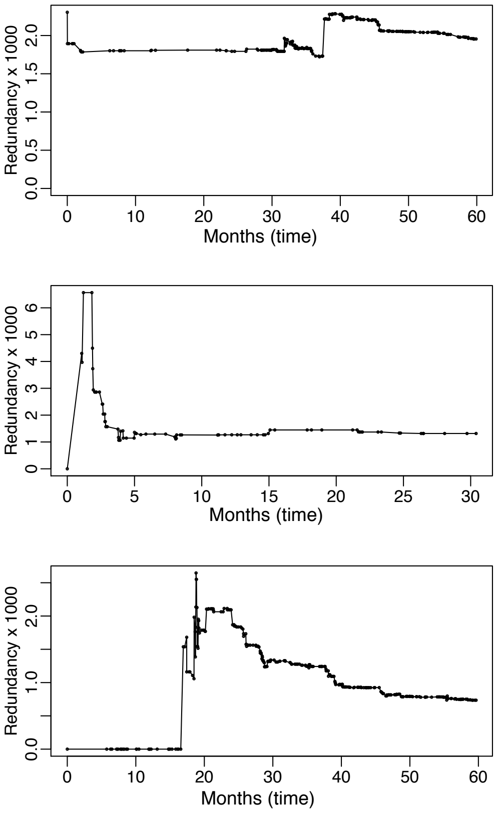
Measure Refinement and Validation. While a high level of similarity between two paragraphs does not always imply the existence of redundant information, in practice, cosine similarity can be used to approximate redundancy.

We validated this measure by asking six volunteers to read and classify the appearance of redundancy in a randomly selected sample of paragraphs. We took a sample of 603 paragraphs (201 for each category). We then randomized the paragraphs into 9 sets, each containing 67 paragraphs for each set, making sure that at least two volunteers examined each text to ensure reliability. We found that a threshold of 0.5 was able to correctly identify redundant paragraphs. In fact, we found that, in 90% of the cases, the text identified by the redundancy algorithm was indeed designated as redundant by the volunteers. For the remaining 10%, the text was identified as redundant because the same quotations were repeated across paragraphs.

3.2 Results

We measured changes in redundancy from when the book was started until its completion, i.e. when it attained ‘featured’ status. (Fig. 1 shows the patterns of redundancy levels over time for three typical books.) We found that none of the 51 featured books contained redundant text when they were completed. However, we did find that books fell into one of three main categories with respect to how redundancy levels changed over the lifetime of the book:

- *Non-Redundant* books (14/51) include negligible amounts of redundancy throughout the book. (Fig. 1.a: The NON-PROGRAMMER'S TUTORIAL FOR PYTHON).
- *Initially Redundant* books (23/51) begin with a great deal of redundancy, followed by either a slow, steady decline or a rapid drop in redundant material. (Fig. 1.b: THIS QUANTUM WORLD).
- *Suddenly Redundant* books (14/51) begin with low levels of redundancy, followed by sharp increases at different points during the development process. The level of redundancy then decreases gradually over time. (Fig. 1.c: C# PROGRAMMING).



a. ***Non-Redundant*:** NON-PROGRAMMER'S TUTORIAL FOR PYTHON

b. ***Initially Redundant*:** THIS QUANTUM WORLD

c. ***Suddenly Redundant*:** C# PROGRAMMING

Fig. 1. Evolution of redundant text over time (in months). Values are not absolute.

We next examined how the above categories affect coordination during the development process. In particular, we examined their correlation with the introduction of redundant text and the perceived quality of each book.

4 Analysis 2: Co-Authorship

With ordinary edited books, a few co-authors divide the work and share top billing. However, in on-line wiki environments, "open-editing" means that potentially thousands of authors may contribute text, with contributions ranging from major sections to just a few words. We are interested in understanding how large numbers of authors who are strangers collaborate to create a 'featured boo'. We thus examine redundancy with respect to other measures, including number of authors, duration of the writing process and length of the book.

Research Questions. We are particularly interested in:

- What is the distribution of authors and their edits in featured books?
- Do the number of authors and the distribution of their edits affect observed redundancy patterns?

4.1 Measures

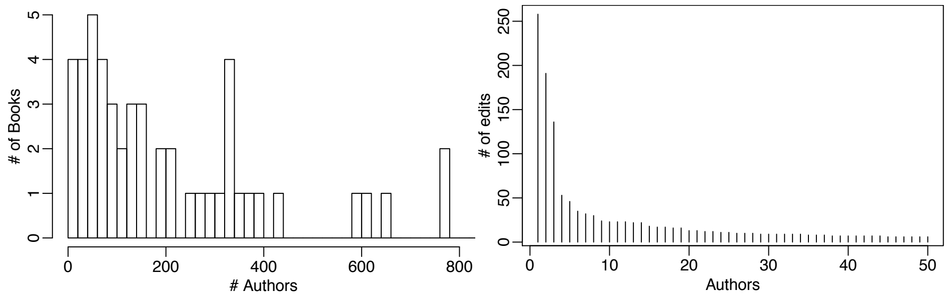
For each book, we collected metadata about its edits: the authors, the number of revisions, and the times when the book was edited. We accessed all versions of the pages within the book, including associated talk pages. We also used the difference between the timestamp on the first revision and the date when the book was classified as featured to calculate its age. We used the three redundancy patterns `RedundancyPatterns` = *Initially-Redundant, Suddenly-Redundant, Non-Redundant* as a factor to analyze these measures.

4.2 Results

We used linear models to examine the possible correlations among continuous measures. Below, we report the p-values for the correlation slopes and the adjusted r^2 that measures the quality of fit. We also conducted one-way ANOVA analyses to examine the effect of the `RedundancyPatterns` on various measures. In turn, these analyses help us to understand the effect of these measures on redundancy patterns.

Authors' Distribution. We can identify different contribution patterns within the histories of different books. The number of contributing authors is highly variable, ranging from 8 to 2631, with a mean of 272±405, a median of 159, a 10% quartile of 34 and a 90% quartile of 614 contributors (see Fig. 2.a).

One common pattern involves a small core of lead authors who write the bulk of the book, aided by a large number of supporting authors who may change only a few words. Panciera et al. [25] found a similar pattern among Wikipedia authors. If we



(a) **(b)**

Fig. 2. a. Distribution of number of authors per book. Not included: 3 books with over 1000 authors (1111, 1039 and 2631). **b.** Distribution of number of edits by authors sorted in descending order (follows a zipf distribution).

rank authors by the number of their edits, the resulting frequency distribution resembles a zipf[5] distribution [36], as in Fig. 2.b.

Next, we consider the number *n* of contributors responsible for x% of the main edits. To compute this, we sort contributors from largest to smallest, based on the number of their edits, as in Fig. 2.b. Here, *n* is the smallest integer *l* for which the first *l* authors made at least x% of the edits. Note that for 57% of the books, one author is responsible for at least 25% of the main edits. The median number of editors responsible for 50% of the main edits is 3.5 and rises to 14 for 75% of the main edits. Interestingly, we did not observe any effect of the `RedundancyPatterns` on the total number of contributors, whether the number of contributors was responsible for 25%, 50% or 75% of the main edits.

For each book, we also computed the Gini coefficient [7], which indicates whether a book was written by a few lead contributors relative to their total number. As above, we did not observe any effect of `RedundancyPatterns` on this coefficient.

The Evolution of Wikibooks. Fig. 3 shows the distribution of the book's age (in days) at the time it became featured (featured time). The age of featured books ranges from 114 to 2034 days with a 1st, 2nd (median) and 3rd quartile of 545, 904 and 1198 days, respectively, with a mean of 900.1±446 days. We observed no effect of the `RedundancyPatterns` on this measure.

Wikibooks contain a varying number of separate pages, ranging from a minimum of 10 pages to a maximum of 646 pages (1st quartile = 26, median = 45, mean = 86±113, 3rd quartile = 90 pages). The number of words in a book ranges from 3945 to 525300 (1st quartile = 3278, median = 4981, mean = 71560±79255, 3rd quartile = 80930 words). We did not observe any effect of the `RedundancyPatterns` on the number of words nor on the number of pages of a book.

When we examined the books, we observed only a marginal correlation between the number contributors (25%, 50%, 75%, and 100% of contributors) and (1) the age,

[5] Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc.

(2) the number of pages, and (3) the number of words. The only significant

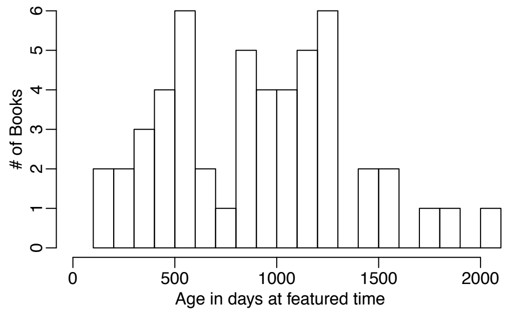


Fig. 3. Distribution of the age of books (in days) when books became featured.

correlation is between the number of contributors and the age of a book at featured time (r^2 = 0.366). More specifically, the age of a book is about 1.2×Num of Contributors + 632 ($p < 0.0001$), after removing the three books with the largest number of authors. This suggests that over time more contributors became involved.

Surprisingly, we found no correlation between the age of a book (at featured time) and its number of words. We did, however, observe a correlation (slightly increasing) between the age of a book and the number of pages ($p = 0.0167$ with a low r^2 = 0.093), and a clear increasing correlation between the number pages and the number of words (r^2 = 0.5945, $p < 0.0001$).

5 Analysis 3: Collaborative Interactions

In traditional collaborations, careful planning is usually essential. However, in on-line collaborations, contributors are under no obligation to follow a plan or discuss issues with each other. Successful collaborative environments such as Wikipedia include both implicit and explicit coordination mechanisms to support collaboration [14]. Wikibooks offer two basic coordination mechanisms: *Predicted pages* specify the structure of the book and identify where additional writing is required. They act as an implicit coordination mechanism in which authors see broken links that point to yet-to-be-written pages and can contribute text accordingly. *Talk pages* enable authors to discuss content and negotiate changes, and act as an explicit coordination mechanism.

Research Questions: We are particularly interested in:

- What is the impact of communication and coordination mechanisms on the appearance of redundancy within the text?
- How does the authors' use of *predicted pages (*broken links) and *talk pages* affect redundancy patterns?

5.1 Measures of Explicit and Implicit coordination

Every Wikibooks page has a talk page in which authors attach comments and discuss content. This enables them to organize the writing task and resolve disputes that may arise and also provided us with a quantitative measure of explicit coordination.

We also created a quantitative measure of implicit coordination based on the structure and number of predicted pages. We also counted the number of edits made by different authors. To differentiate between prolific and casual authors, we sorted authors according to their contributions. We then calculated the inter-quartile range and the Gini coefficient [7] of authors' contributions to measure the skew of contributions within each book. We applied these measures to talk pages and content pages separately.

5.2 Results

Implicit Coordination using Predicted Pages. We found a moderate correlation between the number of predicted pages and the number of pages in a book (r^2 = 0.574). The number of predicted pages represents about 19% ($p < 0.0001$) of the total number of pages. Based on this result, we consider the percentage of predicted pages as a finer measure of implicit coordination.

The `RedundancyPatterns` had a significant effect on the percentage of predicted pages ($p = 0.0101$). Of the total number of books, 14 had no predicted pages, eight belonged to *Initially-Redundant,* three belonged to *Suddenly-Redundant*, and three belonged to *Non-Redundant*. This effect can be observed in the box plot in Fig. 4. A Tukey test shows that the percentage of predicted pages is significantly greater for *Non-Redundant* books.

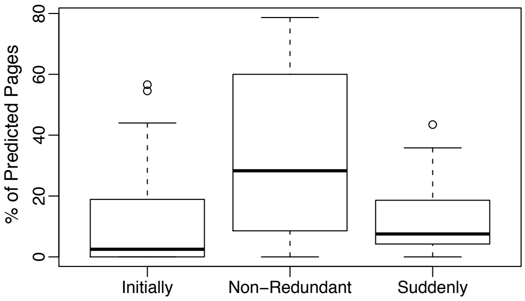


Fig. 4. Box plot showing the percentage of predicted page for each redundancy pattern.

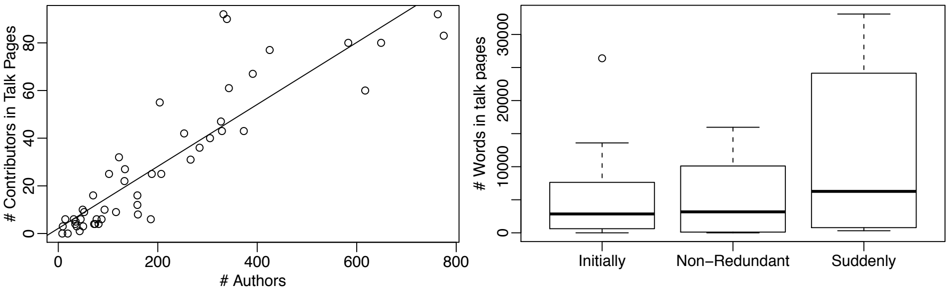
Note that we found no correlation between the percentage of predicted pages and the number of authors (contributions of 25%, 50%, 75%, and 100% of edits) and the number of words in a book. Finally, we only found a marginal decreasing correlation of this measure with the book age (r^2 = 0.06404).

Explicit Coordination using Talk Pages. Not all authors participated in talk page discussions. However, we did find a strong correlation between the total number of authors and how many participated in talk pages (r^2 = 0.788 for books with fewer than

1000 authors (see Fig. 5.a), and $r^2 = 0.923$ for all books). More specifically, about 13% ($p < 0.0001$) of authors participated in talk page discussions.

We found no significant effect of `RedundancyPatterns` on the number of authors into talk pages ($p = 0.079$). However, we found a significant effect of `RedundancyPatterns` on the number of authors for the 25% main talk edits ($p = 0.0325$). In both cases, *Suddenly-Redundant* books have more authors than the two other groups.

We examined the number of words in talk pages as a measure of participation in discussions. We observed that `RedundancyPatterns` has a significant effect on this measure ($p = 0.0171$). *Suddenly-Redundant* books have significantly more words in their talk pages than the other two groups (Fig. 5.b).



(a) **(b)**

Fig. 5. a. Correlation between total number of authors and number of authors who participated in talk pages. **b.** Box plot of the talk page word count for each `RedundancyPatterns`.

We also found a strong correlation between the number of words in talk pages and the time for the book to reach featured status ($p = 0.00123$, $r^2 = 0.177$). Note that we get a similar result (more discussion in *Suddenly-Redundant* books) even when we calculate the number of words in talk pages by day and when we normalize it by the number of pages and words in the book.

Finally, we examined the number of talk page edits. Again, *Suddenly-Redundant* books have significantly more talk page edits and a higher percentage of talk edits relative to main edits. Note that we find no correlation between the percentage of predicted pages and the number of words in talk pages (adjusted $r^2 = -0.003379$, $p = 0.366$). This suggests that these two measures are orthogonal.

6 Summary of Findings

In summary, we found that:

- Redundancy can be categorized according to three distinct patterns:
  1. *Non-redundant* books are well planned from the beginning and require fewer talk pages to reach high-quality status.
  2. *Initially redundant* books begin with high redundancy, which drops as soon as authors use coordination tools to restructure the content.

3. *Suddenly redundant* books display sudden bursts of redundancy that must be resolved, requiring significantly more discussion to reach high-quality status.

- The majority of each book's content (50-75%) is typically written by a small number of lead authors, supported by a larger number of additional authors.
- The *predicted pages* feature acts as an implicit coordination mechanism that does not reduce redundancy in the early phases of a book's development but appears to lower the overall effort of writing a book.
- *Suddenly redundant* books require significantly greater discussion among authors, with a correspondingly high use of talk pages compared to the other two groups, in order to reduce redundancy.
- The number of authors is not correlated with levels or patterns of redundancy, provided that proper coordination mechanisms are used.

7 Revisiting the results through examples

The above analysis showed general trends. This section examines several books in greater detail, including several that are outliers with respect to our statistical analysis.

7.1 Using Implicit Coordination to Avoid Redundancy

We have seen that books with no redundancy tend to have a high percentage of predicted pages. For example, the CONTROL SYSTEMS book has 65 predicted pages, which addresses the majority of the 75 pages in its featured form. The amount of redundancy for this book remains low throughout its history, even though it includes very little discussion (only 1905 out of 74,567 words).

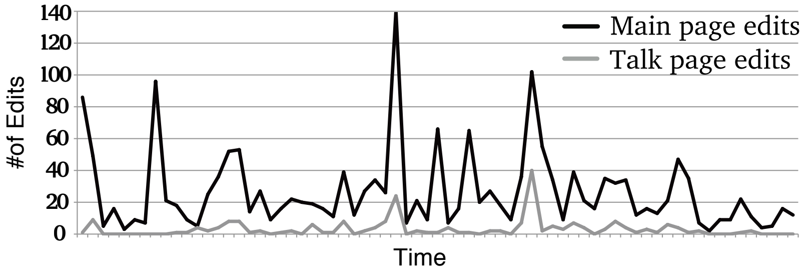


Fig. 6. LATIN book main pages (black) and talk pages (grey) edit history.

The book includes a total of 93 authors with 10 contributors to the talk page discussion. Our analysis suggests that, although these authors engaged in relatively little discussion, they still managed to collaborate without producing redundant text.

Of the books that include no redundancy, not all make extensive use of predicted pages (see Fig. 4). These books achieve coordination through the use of talk page discussions. For example, even though the LATIN book included only two predicted pages of the 87 in its featured form, discussion of the book's structure via talk pages occurred early in its development and continued throughout. Fig. 6 shows the history of main edits and talk page edits and the a strong correlation between them.

The talk page collaboration mechanism allowed 21 lead authors to coordinate the activities of 401 supporting authors on a variety of different topics. This is particularly evident in the talk page in which 77 authors actively discussed the status of the book.

7.2 Using Explicit Coordination to Remove Redundancy

Suddenly redundant books, with their sudden bursts of redundant material, rarely used on-line coordination mechanisms, either via predicted pages or talk pages. Similarly, in *Initially redundant* books, the redundancy curves rise sharply at the beginning due to lack of coordination. The most extreme example is XFORMS, with only one predicted page out of 154. For the first six months, XFORMS was developed mainly by one lead author (Gini coefficient G=0.936), as was SPECIAL RELATIVITY (G=0.874) (3 predicted pages out of 29).

Initial development of ARIMAA involved three lead and twelve supporting authors (G=0.687, 1 predicted page out of 53). In each case, the increase in redundancy is not related to the number of authors who contributed to the books, but rather is due to the authors' approaches to writing, in which they avoided planning the different sections of the book on-line. The three lead authors wrote separately without any interaction among themselves or coordination of future activities. As a result, the other twelve authors who contributed to the book made rather chaotic edits, adding pages that contained redundant information (see also Fig. 1.b).

Lead authors of 17 books compensated for the initial lack of planning by later performing a radical restructuring of the book, with a corresponding reduction in redundancy. For example, the lead author of XFORMS restructured the book into sections, whereas lead authors for both SPECIAL RELATIVITY and ARIMAA used a talk page to direct changes and amendments.

Another example is C# PROGRAMMING, a book in which redundancy increased when four new active authors began to contribute to the book independently, apparently paying little attention to previous contributions by other authors. At this point, the structure was not clear, authors' contributions were not well organized and included no discussion, with only 4 predicted pages and 13 pages of content. Two authors wrote extensively about similar concepts, unaware of each other's contributions.

Finally, one author noticed the overlap and used talk pages to discuss how to restructure the content. This major change in structure was accompanied by a subsequent plunge in redundancy (Fig. 1.c). Over the next months, 50 new authors began to contribute to the book, ultimately leading to an increase in the size of the book by 30%. This demonstrates the importance of structuring books in a logical way, so that authors have a clear idea of what each section should contain, without needing to read the full contents of the book.

7.3 Redundancy might be beneficial

While redundancy within a book as a whole is undesirable, specific increases in redundancy can be beneficial if they lead to restructuring and result in a net decrease in redundancy, e.g. the C# PROGRAMMING example. In some cases the detection of excess redundancy sparks a conversation about the book that generates interest and makes contributing content more appealing. We see this in all ten *Suddenly-Redundant* books and all eight *Initially-Redundant* books.

Another positive restructuring activity was seen in the EUROPEAN HISTORY book, in which a sudden spike in redundancy occurred when a new author contributed text in a new section, even though the same content was already present elsewhere. The other authors chose to retain the new text and re-integrate other relevant text into the new section.

However, sometimes adding text is simply a duplication of effort, as in FORMAL LOGIC in which a casual author increased the redundancy in the book by adding material to a page that was already covered in other pages. Redundancy decreased sharply after this text was removed by the main editor.

8 Discussion

Our analysis of collaboration within Wikibooks is consistent with previous research on Wikipedia with respect to the correlation between communication levels and quality. Like Kittur and Kraut [14], we found correlations between the use of implicit and explicit collaboration strategies, the distribution of authors and article quality. Overall, we found that despite the large number of contributors, most authorship, at least with high-quality 'featured' books, is concentrated among a few lead authors. On average, 75% of a book is written by no more than 14 authors, with a much larger group of supporting contributors who write the rest.

In addition, we observed the effect that collaboration activities have upon the appearance of redundant material within a book. When communication mechanisms are not properly used, authors tend to edit chaotically, which increases the quantity of redundant text. One might expect that the presence of large numbers of authors would lead to duplication of effort and highly redundant text. However, this need not be the case. The number of authors is not correlated with the presence of redundant text. Instead, duplication of effort occurs only when communication mechanisms are used improperly and contributions are chaotic.

Redundancy is normally viewed as a negative characteristic in a book, increasing the effort necessary for the book to become featured. We saw that books with highly redundant content required significantly more coordination effort, even though fewer authors contributed. However, the presence of redundancy can sometimes have a beneficial effect on the quality of the book. When lead authors become aware of redundant text, it triggers the restructuring of redundant pages and a discussion among the authors. In some cases, this can attract new contributors to the book, as seen in the books on C# PROGRAMMING, FORMAL LOGIC and US HISTORY. We do not have a direct measure of how aware authors are of the existence of redundancy, but suggest that lead authors would benefit from tools that draw attention to levels of redundancy since it will encourage them to resolve it, thus improving the quality of the book.

In general, our data suggests if lead authors set the direction, structure and scope of the book from the beginning, a variety of implicit and explicit coordination mechanisms can be used to support subsequent development, aiding future development. As the book matures and coordination requirements ease, tasks may be more effectively distributed to a larger group of authors. This is consistent with other research that suggests that explicit communication through coordination is most beneficial in the early stages of the collaboration, when the structure is unconstrained [14].

Implications for Design: We found that planning of the book's structure, whether implicit (as in predicted pages) or explicit (as in talk pages), has a positive impact upon future coordination of writing activities. Some books were carefully planned from the beginning and could be successfully managed with or without explicit coordination during subsequent writing phases. In contrast, books that were not initially well-planned suffered from a rapid increase in redundancy within the text, either at the beginning, or at different points during the book's development. All of these books eventually ended up as high-quality books, with virtually no redundancy, but such books required significantly more communication and collaboration activities to compensate for the bursts of redundancy.

We believe that authors of large-scale collaborative writing projects will benefit from tools that support the communication and collaboration mechanisms by providing authors with visualizations of the following:

1. *Sections containing redundant text*. This would facilitate restructuring and merging changes, as well as potentially triggering participation by new authors.
2. *Number and types of edits to each page*. This would clarify the structure of each page and let authors focus on writing new sections, rather than extracting redundancy from previously written pages.
3. *Sections containing topics already covered*. This would help authors focus on writing relevant parts of the book instead of duplicating existing content.

9 Summary and Future Work

This paper examines how large groups of volunteer authors coordinated their activities in order to create and edit 51 'featured' or high-quality Wikibooks. We used a combination of information retrieval and statistical methods to develop quantitative measures of coordination activities and identified several factors that are significantly correlated with effective collaboration.

One contribution is the use of redundant text as a measure of coordination effectiveness. Not only did we find that redundancy was inversely correlated with quality, but we also were able to identify different patterns of redundancy over time. *Non-redundant* books are highly coordinated from the very beginning, and progress smoothly to completion. *Initially redundant* books begin with little structure, but once the key authors identify the presence of redundancy and begin to take measures to reduce it, these books also progress smoothly towards completion. *Suddenly redundant* books are poorly planned in the beginning and suffer from spikes in redundancy over time, as people duplicate each other's efforts. These books require by far the most discussion and late-stage coordination in order to become high-quality books.

A second contribution relates to the deeper understanding of the roles of authors in large-scale collaborative writing projects. We found that, even though it appears as though thousands of authors have contributed, in practice, over 75% of the writing is produced by a small core group of lead authors. This implies that these authors need specialized tools for visualizing redundancy, which should help them to more effectively allocate writing tasks and better coordinate everyone's activities.

Our next step will be to modify our algorithms to enable authors to obtain successive snapshots of each level of redundancy. We plan to examine whether providing these authors with ways to visualize this information during the writing process can help to reduce development time and increase the likelihood of writing a high-quality book.

10 Acknowledgments

The authors of this research would like to thank Prof. Hal Abelson for inspiration and invaluable help in setting up this research idea and Theophanis Tsandilas for comments and suggestions on earlier drafts. Ching-Man Au Yeung worked on this research while at the University of Southampton.

References

1. Brooks, F.P.: The Mythical Man-Month: Essays on Software Engineering. Addison-Wesley (1995).
2. Chesney, T.: An empirical examination of Wikipedia's credibility. First Monday 11(11) (2006).
3. Chevalier, F., Dragicevic, P., Bezerianos, A., Fekete, J.D: Using text animated transitions to support navigation in document histories. In: Proc. CHI. pp. 683–692. ACM (2010).
4. Clearwater, S.H., Huberman, B.A., Hogg, T.: Cooperative solution of constraint satisfaction problems. Science 254, 1181–1183 (1991).
5. Emigh, W., Herring, S.C.: Collaborative authoring on the web: A genre analysis of online encyclopedias. In: Proc. HICSS (2005).
6. Encyclopedia Britannica Inc.: Fatally flawed: refuting the recent study on encyclopedic accuracy by the journal Nature (March 2006).
7. Gastwirth, J.L.: The estimation of the Lorenz curve and Gini index. The Review of Economics and Statistics 54(3), 306–316 (1972).
8. Giles, J.: Internet encyclopedia as go head to head. Nature 438,900–901(2005).
9. Goldstein, J., Mittal, V., Carbonell, J., Kantrowitz, M.: Multi-document summarization by sentence extraction. In: Proc. NAACL-ANLP. pp. 40–48. ACL (2000).
10. Gutwin, C., Benford, S., Dyck, J., Fraser, M., Vaghi, I., Greenhalgh, C.: Revealing delay in collaborative environments. In: Proc. CHI. pp. 503–510. ACM (2004).

11. Hill, G.W.: Group versus individual performance: are n+1 heads better than one? Psychological Bulletin 91, 517–539 (1982).
12. Islam, A., Inkpen, D.: Semantic text similarity using corpus-based word similarity and string similarity. ACM Trans. Knowl. Discov. Data 2(2), 1–25 (2008).
13. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with mechanical turk. In: Proc. CHI. pp. 453–456. ACM (2008).
14. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In: Proc. CSCW. pp. 37–46. ACM (2008).
15. Kittur, A., Lee, B., Kraut, R.E.: Coordination in collective intelligence: the role of team structure and task interdependence. In: Proc. CHI. pp. 1495–1504. ACM (2009).
16. Kittur, A., Suh, B., Pendleton, B.A., Chi, E.H.: He says, she says: conflict and coordination in Wikipedia. In: Proc. CHI. pp. 453–462. ACM (2007).
17. Kowalski, G.: Information retrieval systems: theory and implementation. Kluwer Academic (1997).
18. Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse Processes 25, 259–284 (1998).
19. Lerner, J., Pathak, P.A., Tirole, J.: The dynamics of open-source contributors. American Economic Review 96(2), 114–118 (2006).
20. Li, L., Zhou, K., Xue, G.R., Zha, H., Yu, Y.: Enhancing diversity, coverage and balance for summarization through structure learning. In: Proc. WWW. pp. 71–80. ACM (2009).
21. Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., Crockett, K.: Sentence similarity based on semantic nets and corpus statistics. IEEE Trans. Knowl. Data Eng. 18(8), 1138–1150 (2006).
22. Lih, A.: Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In: Proc. ISOJ. pp. 16–17 (2004).
23. Mackay, W.E.: Patterns of sharing customizable software. In: Proc. CSCW. pp. 209–221 ACM (1990).
24. Nardi, B.A., Miller, J.R.: Twinkling lights and nested loops: distributed problem solving and spreadsheet development. Int. J. Man-Mach. Stud. 34(2), 161–184 (1991).
25. Panciera, K., Halfaker, A., Terveen, L.: Wikipedians are born, not made: a study of power editors on wikipedia. In: Proc. GROUP. pp. 51–60. ACM (2009).
26. Pedersen, T., Patwardhan, S.: Wordnet::similarity - measuring the relatedness of concepts. In: Proc. AAAI. pp. 1024–1025 (2004) 27.
27. Raymond, E.S.: The Cathedral and the Bazaar. O'Reilly (2001).
28. Sajjapanroj, S., Bonk, C.J., Lee, M.M., Lin, M.F.: The challenges and successes of wikibookian experts and Wikibook novices: Classroom and community collaborative experiences. In: Proc. AERA (2007).
29. Steiner, I.D.: Group process and productivity. Academic Press (1972).
30. Stewart, G.L.: A meta-analytic review of relationships between team design features and team performance. Journal of Management 32, 26–55 (2006).
31. Thagard, P.: Collaborative knowledge. Nous 31, 242–261 (1997).
32. Turney, P.D.: Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: EMCL. pp. 491–502. Springer-Verlag (2001).
33. Vie ́gas, F.B., Wattenberg, M., Dave, K.: Studying cooperation and conflict between authors with history flow visualizations. In: Proc. CHI. pp. 575–582. ACM (2004).
34. Xiao, Y., Baker, P.B., O'Shea, P.M., Allen, D.W.: Wikibook as college textbook: a case study of college students' participation in writing, editing and using a wikibook as primary course textbook. In: Proc. AERA (2007).
35. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proc. Machine Learning. pp. 412–420 (1997).
36. Zipf, G.K.: The Psychobiology of Language. Houghton-Mifflin (1935).