Studying Cooperation and Conflict between Authors with *history flow* Visualizations

Fernanda B. Viégas
MIT Media Lab
Cambridge, MA 02139 USA
fviegas@media.mit.edu

Martin Wattenberg
IBM Research
Cambridge, MA 02142 USA
mwatten@us.ibm.com

Kushal Dave
IBM Research
Cambridge, MA 02142 USA
kdave@us.ibm.com

ABSTRACT

The Internet has fostered an unconventional and powerful style of collaboration: "wiki" web sites, where every visitor has the power to become an editor. In this paper we investigate the dynamics of Wikipedia, a prominent, thriving wiki. We make three contributions. First, we introduce a new exploratory data analysis tool, the *history flow* visualization, which is effective in revealing patterns within the wiki context and which we believe will be useful in other collaborative situations as well. Second, we discuss several collaboration patterns highlighted by this visualization tool and corroborate them with statistical analysis. Third, we discuss the implications of these patterns for the design and governance of online collaborative social spaces. We focus on the relevance of authorship, the value of community surveillance in ameliorating antisocial behavior, and how authors with competing perspectives negotiate their differences.

Categories & Subject Descriptors

H.5.2: GUI; H.5.3:n Collaborative computing, Web-based interaction.

Keywords

Wiki, revision history, visualization, collaboration, document

INTRODUCTION

Online communities have long allowed people with conflicting perspectives and values to meet and talk—but usually without any need to resolve their differences. Indeed, given the endless arguments often found in traditional online forums, asking that a large group reach consensus online may seem impossible. In recent years, however, new online technologies have arisen that, by their nature, favor consensus building by community members. One example of such a technology is a special kind of web site known as a "wiki." Invented in 1995 by Ward Cunningham [14][9], a defining feature is that any reader of the site may also be an author. Each page has an "edit this page" link at the bottom, allowing users to change the content of the page. This interface supports a higher level of consensus building because a user who disagrees with a statement can very easily delete it. In this sense, the text on wiki pages is content that has survived the critical eye of the community. Since Cunningham's original implementation, wikis have become popular for many purposes both public and private, ranging from knowledge management to education [1][5].

This paper is an examination of the largest public wiki, wikipedia.org (or simply "Wikipedia"), which is a thriving site despite a seemingly unlikely model for success. The founders of Wikipedia wished to create a free online encyclopedia. Rejecting the traditional method of having each article written by an expert and subjected to review, fact-checking and editing, they took the opposite tack: on Wikipedia, content can be added or changed at any time by anyone on the Internet. To many, this approach—so vulnerable to mistakes, ignorance and malice—seems a flatly ridiculous way of producing a serious reference tool. The mystery of Wikipedia is that despite the obvious potential drawbacks of its openness, it has enjoyed significant success. It currently contains articles on more than 100,000 subjects, and from July 2002 to July 2003, it averaged 150,000 page views and 3,300 edits per day [18]. It has attracted many writers, but—more importantly—many readers, suggesting that the articles are worth reading.

In this paper, we describe our investigation into how and why such an open and vulnerable system works. Wikipedia generously makes public its database of articles, along with all past revisions of those articles, providing a rich record of interactions between authors. Mining this vast data set is a challenge: to tackle it, we created a new visualization method, dubbed *history flow*, designed to show relationships between multiple document versions. Exploratory analysis with this visualization revealed complex patterns of cooperation and conflict. We also describe some initial statistical corroboration for the patterns we find. Finally, we propose several hypotheses based on these analyses for how and why this collective authoring environment succeeds.

Our chief conclusion is that Wikipedia and its audience must be viewed as a system in which constant change is a source of strength as well as weakness. The site is subject to frequent vandalism and inaccuracy, just as skeptics might suspect—but the active Wikipedia community rapidly and effectively repairs most damage. Indeed, one type of

CHI 2004, April 24–29, 2004, Vienna, Austria.
Copyright 2004 ACM 1-58113-702-8/04/0004…$5.00

malicious edit we examined is typically repaired within two minutes. Similarly, while rapid content changes mean entries do not have the stability of a print encyclopedia, it also means that entries can take into account relevant news events. We believe such dynamic properties are both interesting in themselves and have implications for the design of other online communities where collaboration and consensus are critical.

Wikipedia history

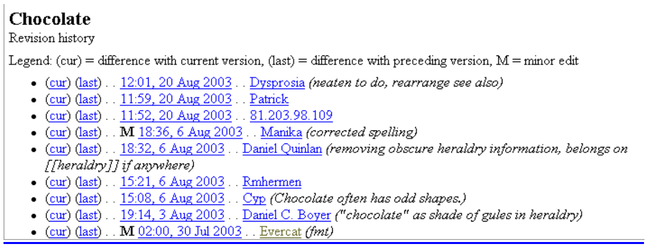
Wikipedia was launched on January 15, 2001. It began as an experimental project related to an earlier encyclopedia site called Nupedia [11]. Nupedia took the conventional approach to encyclopedic writing: articles were written by an expert and approved only after a long review process, fact-checking and editing. Wikipedia instead leveraged the freeform style of interaction developed by Ward Cunningham. While Wikipedia's content grew rapidly, Nupedia's progress has been slow—in the period from October 2001 to April 2003, it released only two new articles [12].

Wiki technology

Wikis rely on server-side technology that allows visitors to make instant updates to a web page via a web interface. Every editable[1] page on a wiki site has an "edit this page" link that visitors can use to alter the contents of the page. Clicking on this link navigates to an editing view with a text field containing the page's contents. The user can edit this text and submit a new version, which will immediately replace the previous one. Editing itself is quite lightweight, using simple markups that are translated into HTML. It is similarly easy to create new pages and new links. In many wikis, including Wikipedia, users have the option of either registering or remaining anonymous. Registered users retain their profile whenever they come back to the site and their changes are logged under their usernames. When anonymous users edit pages, their changes are logged with their IP address.

Most wikis (including Wikipedia) have archiving systems that record all previous edits of a page and make it simple to revert to an earlier version. If the ease of adding a contribution is a distinguishing feature of a wiki, so too, paradoxically, is the ease of removing contributions of others by reverting an edit. This archiving system ensures that no permanent harm can be caused by bad editing.

The archived versions of a page are available to users via a "page history" link. Figure 1 shows a sample page history from Wikipedia. Each row contains: (1) a link to a saved version, (2) a link to the differences between the saved version and the one previous to it, showing what was deleted from and what was inserted to the page, (3) date and time when the change happened, (4) who made the change (in case of an anonymous contributor, the user sees an IP address), and (5) any comments the contributor might have left about the change they made.



Chocolate

Revision history

Legend: (cur) = difference with current version, (last) = difference with preceding version, M = minor edit

- (cur) (last) . . 12:01, 20 Aug 2003 . . Dysprosia *(neaten to do, rearrange see also)*
- (cur) (last) . . 11:59, 20 Aug 2003 . . Patrick
- (cur) (last) . . 11:52, 20 Aug 2003 . . 81.203.98.109
- (cur) (last) . . **M** 18:36, 6 Aug 2003 . . Manika *(corrected spelling)*
- (cur) (last) . . 18:32, 6 Aug 2003 . . Daniel Quinlan *(removing obscure heraldry information, belongs on [[heraldry]] if anywhere)*
- (cur) (last) . . 15:21, 6 Aug 2003 . . Rmhermen
- (cur) (last) . . 15:08, 6 Aug 2003 . . Cyp *(Chocolate often has odd shapes.)*
- (cur) (last) . . 19:14, 3 Aug 2003 . . Daniel C. Boyer *("chocolate" as shade of gules in heraldry)*
- (cur) (last) . . **M** 02:00, 30 Jul 2003 . . Evercat *(fmt)*

Fig 1: *Detail of revision history of Wikipedia's Chocolate page.*

Finally, wikis have a "recent changes" page that lists the latest edits that have taken place across the site. This is one important way in which users of a wiki track activity since their last visit.

Wikipedia enhancements

Some critical features in Wikipedia are incidental or even absent in other wiki implementations. Wikipedia allows users to keep a "watch list" of pages they wish to monitor closely. When a page in someone's watch list is modified, the user is notified via email. This is an effective means for topic experts and serious Wikipedians to scrutinize changes made to specific pages and fix acts of vandalism such as mass deletions. Watch lists function as alerting mechanisms for wiki communities.

The Wikipedia community also sets up secondary pages that are devoted to the discussion of issues surrounding the topics on "real" pages; these are sometimes called "talk pages." They represent an attempt to separate what is "real" information from discussions about what should and should not be on the real page.

THE *HISTORY FLOW* VISUALIZATION TECHNIQUE

Historical information on how communal documents are created and edited is critical for understanding collaborative dynamics within communities [13]. Wikipedia makes its entire database of version histories available for download, a boon to researchers. Making sense of the history for even a single entry, however, is not straightforward. The sheer number of versions can be daunting: as of August 2003, the entry for Microsoft had 198 versions comprising 6.2 MB of text; to get an idea of how much information this is, imagine a table like the one in Fig. 1 but 22 times larger. Moreover, significant information is often not contained in individual versions, but in the differences in the text of an entry from one version to the next. Such differences highlight editing choices, emphasizing what does and does not survive over time.

[1] Wikis can also have pages that are protected and, therefore, only editable by administrators. Sometimes wikis will protect their Front Page; Wikipedia, for instance, does that.

Wikipedia provides a method of viewing differences, similar to that found in source control systems such Visual Source Safe [17]. This interface suffers from two drawbacks: First, it only shows differences between two versions at once. Second, it records differences only on a paragraph level (a change in a comma might cause a two-page paragraph to be marked as deleted). Both problems made examination of version histories extremely cumbersome. Since no commercial tools were available that solved both problems, we created a new technique, a simple but effective visualization tool, dubbed *history flow.*

The goal of *history flow* is to make broad trends in revision histories immediately visible, while preserving details for closer examination. We found this method invaluable in analyzing the Wikipedia data set, but we believe it is of independent interest and may be applicable in many other collaborative situations. One particularly promising avenue is investigating patterns in large-scale software development.

To explain the technique, we consider a hypothetical scenario where three people—Mary, Suzanne, and Andrew —collaborate in writing a document. Each version of the document is represented by a vertical "revision line" with length proportional to the length of its text. The contributors are each assigned a different color in the visualization, and sections of each revision line are colored according to who originally authored them [Fig. 2A].

In our scenario Mary creates the page and thus the first revision line [Fig. 2A, at left] is entirely black, Mary's author color. Now imagine that Suzanne adds text to the end of what Mary wrote. In the revision line for the second version [second line from left, Fig. 2A], this addition shows up in Suzanne's author color as an appended line at the bottom of Mary's original line. The overall length of the document grows in the second version. On "version 3" Andrew deletes a portion of Mary's original text and introduces a small contribution between Mary's and Suzanne's texts. Finally, in "version 4" Suzanne inserts some text towards the top of the page, in the middle of what has survived of Mary's original text [Fig. 2A, right].

The sequence of revision lines shown in Fig. 2A makes up the skeleton of the visualization, but these lines alone omit critical information. In particular, it is hard to see how the different versions relate. The key step in a *history flow* diagram is to visually link sections of text that have been kept the same between consecutive versions. To do so, we draw shaded connections between corresponding segments on adjacent revision lines [Fig. 2B]. Pieces of text that do not have correspondence in the next (or previous) version are not connected and the user sees a resulting gap in the visualization, clearly highlighting deletions and insertions.

One helpful variation on the *history flow* method is to use the spacing of revision lines to indicate the passage of time. Instead of the regular spacing shown in Figs. 2A and 2B, we let the space between successive revision lines be proportional to the time between the revision dates [Fig. 2C]. This alternative view which we call *space by dat*e, de-emphasizes revisions that come in rapid succession and, as discussed later, can be quite revealing of the rhythms of collaboration among authors.

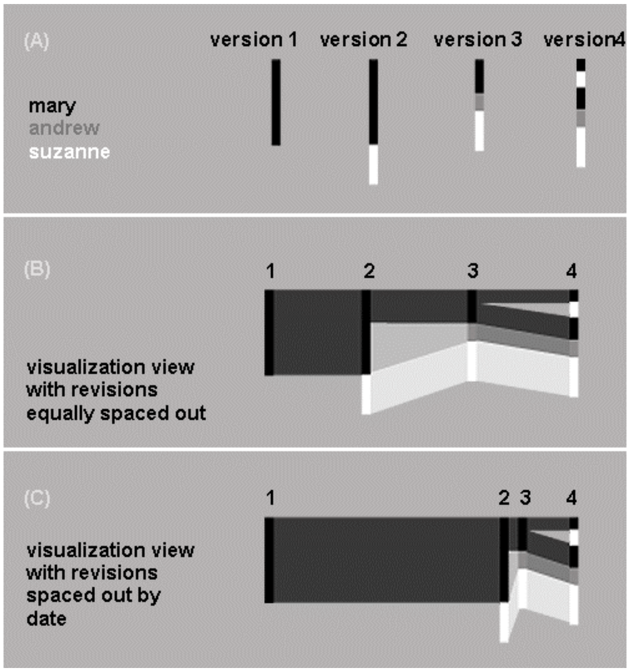


Fig 2: *explanation of history flow's visualization mechanism*

When applied to complex version histories, *history flow* can produce striking results. Figure 3, for example, shows a view of the version history for the Wikipedia entry for Microsoft.

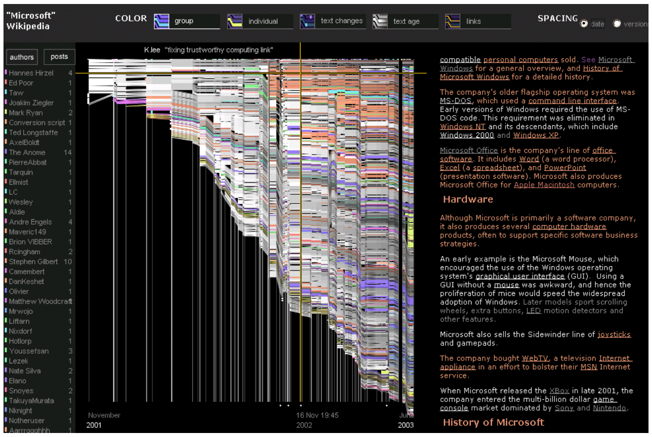
User interface

The interface of the visualization tool is relatively simple. The bulk of the screen is devoted to the *history flow* visualization itself [Fig. 3]. Above it are buttons that let the user change the color scheme in the visualization, for example, highlighting only contributions by a given author. To the side of it is a text panel closely linked with the visualization, so that if the user moves a set of crosshairs to a location on the visualization, the text view shows the text for the corresponding version and position within that version. Conversely, scrolling the text view will move the marker on the visualization. This tight linking of overview and detail was critical for effective analysis.

When the user selects a revision line, we provide additional annotations to help understand its context. The author's comment is displayed at the top of the revision line, and the date of the selected version (down to the nearest minute) is displayed at the bottom. Additionally, all other versions by that author are highlighted.

Implementation notes and related work

Finding matching sections of two document revisions is a well-studied problem in computer science, with many



authors	posts
Hannes Hirzel	4
Ed Poor	1
Taw	1
Joakim Ziegler	1
Mark Ryan	2
Conversion script	1
Ted Longstaffe	1
AxelBoldt	1
The Anome	14
PierreAbbat	1
Tarquin	1
Ellmist	1
LC	1
Wesley	1
Aldie	1
Andre Engels	4
Maveric149	1
Brion VIBBER	1
Rcingham	2
Stephen Gilbert	10
Camembert	1
DanKeshet	1
Olivier	1
Matthew Woodcraft	[illegible]
Mrwojo	1
Liftarn	1
Nixdorf	1
Hotlorp	1
Youssefsan	3
Lezek	1
Nate Silva	2
Elano	1
Snoyes	2
TakuyaMurata	1
Nknight	1
Notheruser	1
Aarrrggghhh	1

compatible personal computers sold. See Microsoft Windows for a general overview, and History of Microsoft Windows for a detailed history.

The company's older flagship operating system was MS-DOS, which used a command line interface. Early versions of Windows required the use of MS-DOS code. This requirement was eliminated in Windows NT and its descendants, which include Windows 2000 and Windows XP.

Microsoft Office is the company's line of office software. It includes Word (a word processor), Excel (a spreadsheet), and PowerPoint (presentation software). Microsoft also produces Microsoft Office for Apple Macintosh computers.

Hardware

Although Microsoft is primarily a software company, it also produces several computer hardware products, often to support specific software business strategies.

An early example is the Microsoft Mouse, which encouraged the use of the Windows operating system's graphical user interface (GUI). Using a GUI without a mouse was awkward, and hence the proliferation of mice would speed the widespread adoption of Windows. Later models sport scrolling wheels, extra buttons, LED motion detectors and other features.

Microsoft also sells the Sidewinder line of joysticks and gamepads.

The company bought WebTV, a television Internet appliance in an effort to bolster their MSN Internet service.

When Microsoft released the XBox in late 2001, the company entered the multi-billion dollar game console market dominated by Sony and Nintendo.

History of Microsoft

Fig 3: *history flow user interface showing the Microsoft page on Wikipedia; on the right we see the contents of the page, on the left we see all the authors who have contributed to this page; the center panel shows the visualization*

possible solutions. For *history flow*, we chose a simple technique that works by matching up tokens [7] —in our program "sentences" are defined as pieces of text delimited by periods or html tags—which gives decent results with sufficient efficiency. One problem with this approach is that tiny changes, such as the addition of a single comma, will show up as a change to an entire sentence, but even this level of granularity is a large improvement over the paragraph-level view that is the Wikipedia default.

There are many existing methods for visualizing document revisions. Several popular source control interfaces can color-code changed regions in files and show a side-by-side comparison of two files, graphically connecting matching sections [17]. Other methods use a thumbnail view of a program, with line-by-line coloring to indicate authorship or age [4]. There is also artwork depicting software code histories that displays differences between multiple versions [3]. Visually, *history flow* diagrams have some similarity to Theme River [6] and to parallel coordinates systems [8], but our method depicts a completely different type of data and, our vertical axes function differently.

PATTERNS OF COOPERATION AND CONFLICT

We used the *history flow* method to examine in detail more than 70 different Wikipedia page histories. Our examination revealed several common patterns of collaboration and negotiation. These patterns represent some of the techniques that this community has developed to deal with antisocial behavior, disputes, and the determination of what is off topic on a page. We now describe several of these patterns: vandalism and repair; anonymity versus named authorship; negotiation; and content stability.

STATISTICAL ANALYSIS: METHOD AND DATA

The *history flow* visualization revealed some fascinating patterns, but examining pages by hand has obvious limitations. To find additional evidence for the patterns that we spotted, we relied on large-scale statistical analysis of the Wikipedia archives [16]. The statistics in the sections below were derived from data that represents the state of the encyclopedia's history as of May 2003. To derive statistics, the archives were loaded into a MySQL database and queries were made using standard SQL syntax. The database contains both "content pages" which represent entries in the encyclopedia as well as "talk pages", which represent discussion about the encyclopedia itself. Unless otherwise specified, statistics we cite below are from the set of content pages only. There were 130,596 such pages, with an average of 5.7 versions for each. 79,813 content pages had been revised at least once.

Vandalism and repair

Wikis are vulnerable to malicious edits or "vandalism," which can take a surprising array of forms. The true scope of vandalism became clear to us upon viewing the *history flow* visualizations.

Mass deletions —edits that remove most of the contents of a page—constitute one common form of vandalism in Wikipedia, and are easily spotted in our visualizations because they appear as breaks in the continuous horizontal flow of changes. In the *history flow* diagram for the Wikipedia page on Abortion [Fig. 4], the abrupt black gutters represent instances of mass deletions. Fig. 4 is a view that equally spaces out revisions. When, however, we look at the same data spaced by date [Fig. 5], we notice that there are no interruptions. The instances of mass deletion were fixed so quickly that they cannot be seen when revisions are spaced by date. This pattern appeared in almost every instance of a vandalized page that we examined by hand. Many of the pages we examined that had long revision histories (more than 50 versions) had suffered at least one act of vandalism.

In some cases the Wikipedia community itself cannot agree on whether an edit constitutes vandalism or not. In fact there is a vandalism-tracking page where users discuss and coordinate responses to specific instances of vandalism.

Because of their short-lived nature in the Wikipedia site, damaging acts often appear in *history flow* visualizations as single-version perturbations of the bigger, general flow of a page's evolutionary history: either one-version deletions or one-version insertions of content.

The variety of vandalism found in Wikipedia can be astounding; five common types are listed below:

1. Mass deletion deletion of all contents on a page.

2. Offensive copy: insertion of vulgarities or slurs.

3. Phony copy: insertion of text unrelated to the page topic. E.g. on the Chemistry page, a user inserted the full text from the "Windows 98 readme" file.

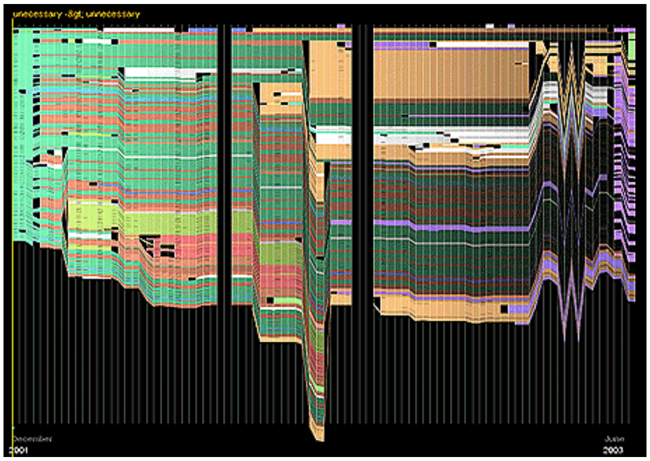


Fig 4: *history flow for "Abortion" page, versions equally spaced.*

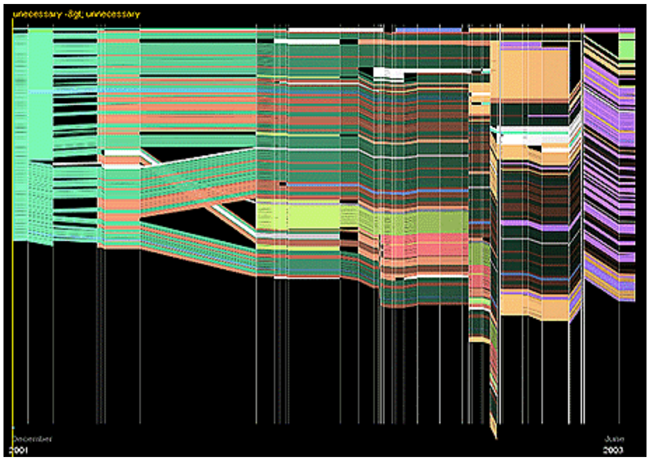


Fig 5: *history flow for "Abortion" page, spaced by date*

4. Phony redirection: Often pages contain only a redirect link to a more precise term (e.g. "IBM" redirects to "International Business Machines."), but redirects can also be malicious., linking to an unrelated or offensive term. "Israel" was at one point redirected to "feces." Note that a phony redirect implies familiarity with Wikipedia's editing mechanisms.

5. Idiosyncratic copy: adding text that is related to the topic of the page but which is clearly one-sided, not of general interest, or inflammatory; these may be long pieces of text. Examples range from "Islam" where a visitor pasted long prayer passages from the Koran, to "Cat" where a reader posted a lengthy diatribe on the Unix *cat* command.

Statistical corroboration

We sought statistical corroboration for our impression that vandalism is frequent and that it is fixed very rapidly. It is essentially impossible to find a crisp definition of vandalism —as mentioned above, the Wikipedia community argues about it frequently—but there are certain computable markers that indicate vandalism.

We defined a mass deletion ("Mass delete," or MD, in Table 1) to be a version that was at least 90% smaller than the previous maximum size for the page, did not redirect the user to a different page, and wasn't created by a Wikipedia administrator. While this category included many malicious edits, it also included many edits that, on close inspection, seemed well intentioned. To pinpoint a group of purely ill-intentioned edits, we looked at mass deletions where the remaining text included the word "fuck," [2] labeled "MD obscene" in Table 1. This group included 47 edits, all of which seemed (to the authors of this paper) unmistakably malicious.

We then looked at the *survival time*, that is, the total time that these edits remained on the site. Time on site is strongly skewed positive, so we computed both mean and median times. The results provide corroboration for the conclusions drawn from the visualizations. It is especially dramatic that half of mass deletions are modified within 3 minutes, and half of vulgar mass deletions are modified within 2 minutes.

Revision Type	**Number**	**Mean time**	**Median time**
All content	618,502	22.3 days	90.4 minutes
Mass delete (MD)	3,574	7.7 days	2.8 minutes
MD obscene	47	1.8 days	1.7 minutes

Table 1: *Survival time for different kinds of revisions.*

Negotiation

A second pattern revealed by our visualizations is a zigzag arrangement that lasts for a few versions before dying out [Fig. 6]. On closer inspection we realized these patterns indicated what the Wikipedia community calls "edit wars," interactions where two people or groups alternate between versions of the page. Some edit wars last as long as 20 consecutive versions.

To our surprise we found that edit wars are not confined to controversial topics. One such example is the page on Chocolate [Fig. 6]: two users fought over whether a kind of chocolate sculpture called "coulage" really existed and consequently, whether or not the paragraph about it should appear on the page. This discussion resulted in 12 consecutive versions of reverting back and forth between two versions. Eventually the paragraph was taken out for good.

Our investigation shows that conflict can take several forms and can occur in different forums. One forum where people preemptively try to resolve disagreements is via their comments on why they edited something on a page. While using *history flow*, we noticed that comments on consecutive revisions often read as a conversation between authors, rather than a mere summary of edits. Frequently authors preemptively address possible objections or direct questions to each other.

[2] This particular obscene word was chosen based on its disproportionately frequent use in acts of vandalism. An explanation of vandals' attraction to this specific obscenity is beyond the scope of this paper.

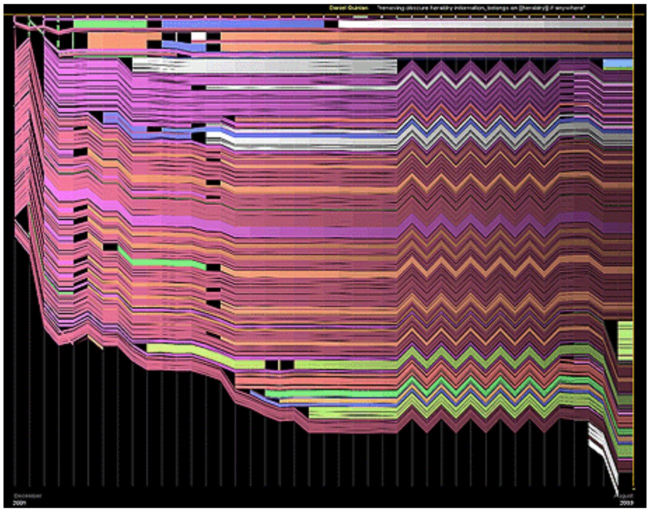


Fig 6: "Chocolate" page spaced out by number of versions; we can see the zigzag pattern of an edit war.

The talk pages that accompany each Wikipedia entry are explicitly designed for resolving disputes, and are frequently used for that purpose. The talk pages function as extensions of edit comments, but afford more room for people to argue their positions. Our observations suggest that when people cannot convince others of why their edits are valid via the comments they leave, the discussion escalates into the talk pages. Talk pages comprise a significant amount of the content on Wikipedia; the May 2003 database snapshot contains more than 11,000 "meta" pages, accounting for 17% of all versions in the May 2003 database.

Authorship

Explicit authorship of contributions on wiki pages is an issue of some contention among wiki users; whereas some feel that authorship is an important part of social collaboration in the sense that it adds context to interactions, others feel that authorship data is irrelevant and sometimes even detrimental to the creation of truly communal repositories of knowledge [20].

An explicit goal of Wikipedia is to create encyclopedic entries that are "neutral" instead of expressing personal biases. This "neutral point of view" ("NPOV,"in Wikipedia shorthand) is a touchstone of the Wikipedia community, frequently referred to in comments and talk pages. One reflection of the NPOV policy is that contributions to article pages are not signed within the page itself. However, on the discussion-oriented talk pages that accompany articles, most authors sign their comments. This kind of conversation page makes for a different social space from the regular Wikipedia article page. It is an important social environment where conflict can develop and settle more naturally.

A small sample of frequent Wikipedia users said that they rely on authorship information when browsing the RecentChanges page or the history page of a specific Wikipedia article. These page "watchers" become familiar with the names of regular contributors to the pages they watch and are constantly on the lookout for any unfamiliar names and unfamiliar IP addresses (the "signature" left by anonymous contributors). First-time contributors represent a potential threat of vandalism and therefore their edits are closely scrutinized. On the other hand, there is also the possibility that a newcomer is someone who may be unfamiliar with Wikipedia standards. In either case the article merits a second look.

Another pattern related to authorship and easily identifiable in *history flow* is the variation in the level of anonymous contributions across different pages. There is huge inconsistency between individual pages in the proportion of anonymous contributions over time. Roughly 31% of the versions in the May 2003 database were contributed by anonymous authors. Some pages have been largely written by anonymous contributors (in our visualization, these show up as diagrams mostly in shades of gray). Examples of such pages include: Microsoft [Fig. 3], Sex, Music, Libertarianism, Creation, and Computer. Other pages have few anonymous contributions ever in their history, for example: Mythology, Evolution, Design, and Brazil [Fig. 7]. We have not observed a clear preference either on the side of the anonymous users or otherwise for specific topics or clusters of topics. More in-depth analysis is needed to help determine what can account for this distinction.

There is also no clear connection between anonymity and vandalism. We observed instances of vandalism by users with (sometimes tauntingly offensive) registered usernames. Conversely, there are users that contribute quite a lot to the site but who choose to remain anonymous. We found one particular case where an anonymous contributor to the page on Capitalism edited the page 55 times between Nov. 22, 2002 and Jun. 26, 2003. This person's contributions were quite substantial and were kept by subsequent contributors.

Temporal patterns and content stability

A *history flow* visualization is, in effect, a fancy graph of how the length of a page varies over time—and it turns out that even this simple measure holds some surprises. One might guess that pages would tend to stabilize over time. The visualization tells another story. Most pages we viewed showed continual change in size and turnover in text. As examples, figure 3 (Microsoft) shows an instance of near-constant growth; figure 4 (Abortion) shows an example of growth and shrinkage. Note that shrinkage can occur either when copy is deleted or when a large section of the page is redirected to a new site (for instance, the most dramatic shrinkage in the Abortion page in figure 5 is due to material being shifted to an entry on abortion in Ireland.)

Our inspection of visualizations suggested several other patterns which deserve mention. One pattern we call *first-mover advantage*. The initial text of a page tends to survive

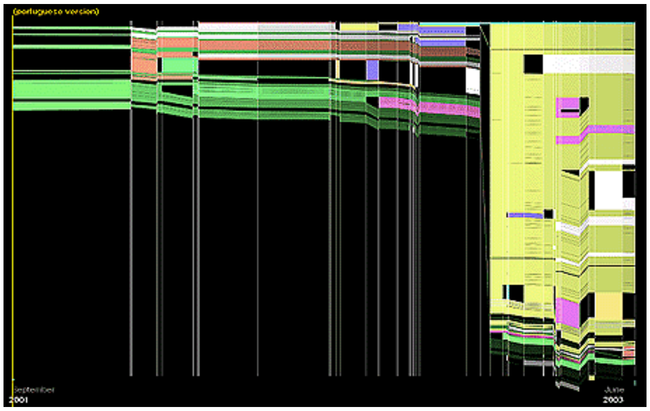


Fig 7: *"Brazil" page showing abrupt growth and few anonymous contributions.*

longer and tends to suffer fewer modifications than later contributions to the same page. Our hypothesis is that the first person to create a page generally sets the tone of the article on that page and, therefore, their text usually has the highest survival rate.

A second pattern is that people tend to delete and insert text more frequently than moving text in an article. In other words, we see many more "gaps" in the visualization than the type of crossing lines in Fig. 5. One explanation may be that the editing window of Wikipedia pages is by default 25 lines long, making it hard for one to see long articles in their entirety. Consequently, the task of moving things around becomes a lot more cumbersome than if one could access the entire text at once. If correct, this explanation could help guide wiki developers in building more user-friendly editors for wiki pages.

Statistical corroboration

We wished to directly measure the level of instability of Wikipedia pages, but obtaining meaningful numbers for stability is difficult for two reasons. First, it would take a prohibitive amount of time to run a fine-grained differencing algorithm on hundreds of thousands of versions, especially one able to distinguish accurately between a change of an entire sentence and an addition of a single comma. Second, and more seriously, Wikipedia has existed for a short time, during which the number of readers (hence editors) has grown tremendously, thus making time-based measures hard to interpret.

We therefore focused on size change as a simple measure of change in content. Using several measures, we found little evidence for stability. For instance, there are 273 pages on Wikipedia which had more than 100 versions as of May 2003. A graph plotting average version size in this subset versus version number [Fig. 8] shows steady growth. Thus, as suggested by the *history flow* visualization, heavily edited pages seem not to converge in size. To take another example, 21% of edits reduced the size of a page, with 6% decreasing it by more than 50 characters. Such cuts can be beneficial (tightening prose, eliminating irrelevant information) but at the same time they make citing Wikipedia as a source problematic, since the information cited may be removed from the page. Rapid turnover also means that news events may be assimilated with a speed that is impossible in a print encyclopedia. Within a week of the U.S. invasion of Iraq in 2003, for example, a page devoted entirely to that topic had been written, and the entry on Iraq itself tripled in size in the weeks after the invasion began.

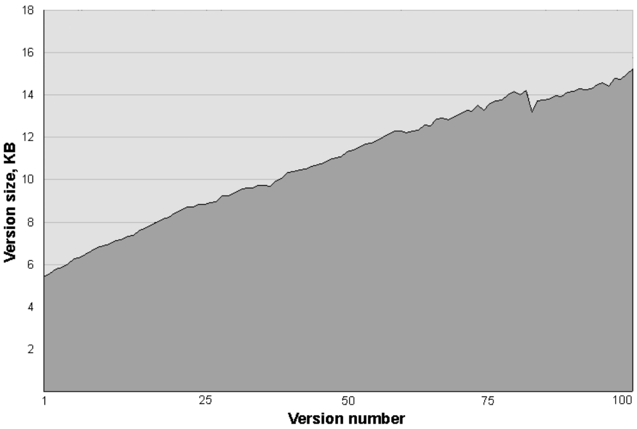


Fig 8: *Graph of version number versus average version size (in kilobytes) shows steady growth for pages with at least 100 edits.*

DISCUSSION

The patterns described above show that Wikipedia has enjoyed significant success as a community in which people with disparate perspectives can collaborate to create a single document. A key question for designers of online communities is: How did they do it? In other words, what design decisions allow Wikipedia to create the social structures that make it a successful system? A full answer to this question is beyond the scope of this paper but is an important line of investigation. Here we propose three hypotheses that may explain Wikipedia's success, and that may be useful as a starting point for future research. The common thread in these hypotheses is that Wikipedia encourages community introspection: that is, it is strongly designed so that members watch each other, talk about each other's contributions, and directly address the fact that they must reach consensus.

First, the watchlists provide a mechanism for community surveillance, and may be responsible for the extremely rapid response to vandalism noted above. Second, the talk pages and other non-content spaces help in removing "meta-level" discussions from the main encyclopedia. Indeed, the May 2003 database snapshot contains more than 11,000 talk pages, a large amount of discussion. Yet it is extremely rare to find discussion about an article embedded in the article itself. Finally, the group consensus that a "neutral point of view" is to be desired provides both common ground and rough guidelines for resolving disputes.

FUTURE WORK

The work here suggests several directions for future investigation. Our statistical analysis is preliminary and can be extended in many directions, for example better algorithmic detection of vandalism. Another area that bears further investigation is the relationship that talk pages have to the article pages in Wikipedia and how the discussions there compare to the collaboration patterns in the articles. It may be revealing, also, to compare the Wikipedia versions for different languages and look for cultural differences. Finally, it would be interesting to investigate other wiki sites. New patterns may emerge, and comparisons with Wikipedia should make general patterns clear. Wikipedia is much larger (with a much more diverse readership) than most wikis and it would be interesting to see differences in collaboration in other types of groups.

CONCLUSION

When visiting a wiki, one is greeted with what looks like a conventional static Web site. Yet this serene façade conceals a more agitated reality of constant communal editing. Hundreds, sometimes thousands of busy hands insert words, create new pages, delete paragraphs, manicure the contents of the site.

To better understand the ebb and flow of this editing frenzy we have introduced *history flow*, a tool for visualizing how collaborative documents evolve over time. This technique reveals some of the patterns that have emerged within Wikipedia: its surprisingly effective self-healing capabilities, the variety of negotiation processes used in reaching consensus; the diversity of authorship, the bursty rhythms of page editing, and the constant change in page contents. In turn, these facts point to some of the key social mechanisms of the community: the importance of having forums for resolving conflicts and the value of fast, efficient notification of changes to aid surveillance.

Without the aid of *history flow*, it would have been a daunting task to piece together the collaboration patterns described here. The efficacy of *history flow* in highlighting patterns of behavior suggests that visualization is a technique well-suited to records of social behavior. One speculation is that social interaction is often characterized by mostly normal behavior punctuated by outlying abnormal episodes, and information visualization can be an excellent way to simultaneously show broad trends and outlying data points.

We believe that the results described in this paper are of general interest for several reasons. First, Wikipedia is just one of many wiki sites that make no distinction between readers and writers. We believe our findings have relevance for the design of other wiki sites, especially as they scale up in size. Second, the *history flow* visualization method can be utilized in other situations that involve heavily revised documents by multiple authors such as software version control systems for instance. Finally, the ability to better understand the mechanisms for reaching consensus described here may apply in other contexts and the "self-healing" qualities that Wikipedia promotes may turn out to be a general principle of long-lived online communities.

ACKNOWLEDGMENTS

We would like to thank the Wikipedia community and Jimmy Wales who were gracious enough to provide us with data and answer questions. We would also like to thank our colleagues at IBM and Clay Shirky for their input and insightful discussions.

REFERENCES

1. Aronsson, Lars. *Operation of a Large Scale, General Purpose Wiki Website: Experience from susning.nu's first nine months in service*. In proceedings of the International ICCC/IFIP Conference on Electronic Publishing, 2002.
2. Dieberger, A. and Guzdial, M. *CoWeb – Experiences with Collaborative Web Spaces*. In From Usenet to CoWebs: Interacting with Social Information Spaces. Springer Verlag, 2002.
3. Fry, B. revisionist http://acg.media.mit.edu/people/fry/revisionist
4. Baker, M. J., Eick S. G., *Space Filling Software Visualization*. Journal of Visual Languages and Computing, Vol. 6, pp 119-133, 1995.
5. Guzdial, M., Rick, J., Kerimbaev, B. (2000) "Recognizing and Supporting Roles in CSCW" Proceedings ACM CSCW 2000.
6. Havre, S., Hetzler, B., and Nowell, L., *ThemeRiver^{TM}: In Search of Trends, Patterns, and Relationships*. IEEE Transactions on Visualization and Computer Graphics. 8(1):9-20; 2002.
7. Heckel, Paul. *A Technique for Isolating Differences Between Files*. Communications of the ACM 21(4), pp. 264—268, April 1978.
8. Inselberg, A., *The plane with parallel coordinates*. The Visual Computer, 1(2):69--92, 1985.
9. Leuf, B., Cunningham, W. The Wiki Way. Addison-Wesley, 2001.
10. meta.wikipedia.org
11. Nupedia site: http://www.nupedia.com/
12. http://www.wikipedia.org/wiki/Nupedia
13. Smith, M., Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of USENET. In Communities in Cyberspace, Routledge Press, 1999.
14. Wiki Wiki Web: http://c2.com/cgi/wiki?WikiWikiWeb.
15. Wikipedia site: http://www.wikipedia.org/
16. Wikipedia database page: http://download.wikipedia.org/
17. Visual Source Safe (Microsoft, http://msdn.microsoft.com/ssafe/)
18. wikistat: http://www.wikipedia.org/wiki/Wikipedia:Statistics
19. http://meta.wikipedia.org/wiki/History_of_Wikipedia
20. http://c2.com/cgi/wiki?ThreadModeConsideredHarmful