Talk Before You Type: Coordination in Wikipedia

Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, Frank van Ham
Visual Communication Lab, IBM Research
{viegasf, mwatten, jesse.kriss, fvanham}@us.ibm.com

Abstract

Wikipedia, the online encyclopedia, has attracted attention both because of its popularity and its unconventional policy of letting anyone on the internet edit its articles. This paper describes the results of an empirical analysis of Wikipedia and discusses ways in which the Wikipedia community has evolved as it has grown. We contrast our findings with an earlier study [11] and present three main results. First, the community maintains a strong resilience to malicious editing, despite tremendous growth and high traffic. Second, the fastest growing areas of Wikipedia are devoted to coordination and organization. Finally, we focus on a particular set of pages used to coordinate work, the "Talk" pages. By manually coding the content of a subset of these pages, we find that these pages serve many purposes, notably supporting strategic planning of edits and enforcement of standard guidelines and conventions. Our results suggest that despite the potential for anarchy, the Wikipedia community places a strong emphasis on group coordination, policy, and process.

1. Introduction

Wikipedia, the online encyclopedia that anyone can edit, has gone from being a Web curiosity to becoming a powerful source of information for both the online and offline worlds. It is not uncommon to see Wikipedia pages being used as references in media news stories, and students sometimes turn to Wikipedia as a source of learning materials. Such extensive usage, and the implications of Wikipedia as an authoritative source have become hotly debated topics in academia. Researchers have discussed the implications of having such an open-ended forum serve as a trustworthy source of information [3][8][10]. Others have hailed the power of Wikipedia to amass astounding quantities of knowledge. Yochai Benkler suggests that Wikipedia is typical of a new "commons-based" mode of economic production [1].

Despite this far-reaching discussion, at this point little empirical data has been published about the fundamental inner workings of Wikipedia. What is the structure, or "shape," of this sprawling site? How big is it and how much has it grown in recent years? How many people contribute to Wikipedia? How many times is the average page edited? Have contributions changed over the years? By exploring these questions, this study reveals some of the basic collaboration mechanisms present in Wikipedia.

We investigate how Wikipedia has evolved in recent years by repeating a series of enquiries and visualizations done by Viégas, Wattenberg and Dave in 2003 [11]. We compare and contrast our results to those of [11] and note that administrative and coordinating elements seem to be growing at a faster pace than the bulk of articles in the encyclopedia. One hypothesis, therefore, is that Wikipedia is becoming less anarchic and more driven by policies and guidelines. In order to examine this phenomenon more closely, we have turned our attention to so-called "Talk" pages in Wikipedia, where much of the coordination work occurs. Our results indicate that Talk pages serve a variety of important functions in the maintenance of articles, ranging from strategic planning of editing activities to the enforcement of Wikipedia policies and conduct guidelines.

2. Related Work

Wikipedia's popularity and the subject of its reliability have attracted the attention of academics and the media alike. In a study that compared Wikipedia and Everything2 articles on the same topics, Emigh and Herring found that Wikipedia entries are stylistically similar to traditional, printed sources such as the expert-created Columbia Encyclopedia, in terms of formality and language standardization [3]. They attribute this phenomenon to the high degree of post-production editorial control afforded by Wikipedia—for instance, the ability to easily edit other's entries. In a study carried out by Nature, the level of accuracy in Wikipedia entries was found to be the same as that of Britannica articles [5]. Stvilia et al. investigated how the Wikipedia community establishes and improves information quality through discussions in Talk pages [10]. After analyzing the contents of a series of Talk pages—in a similar fashion to the present study— they found that these pages play a crucial role in letting

1530-1605/07 $20.00 ©2007 IEEE

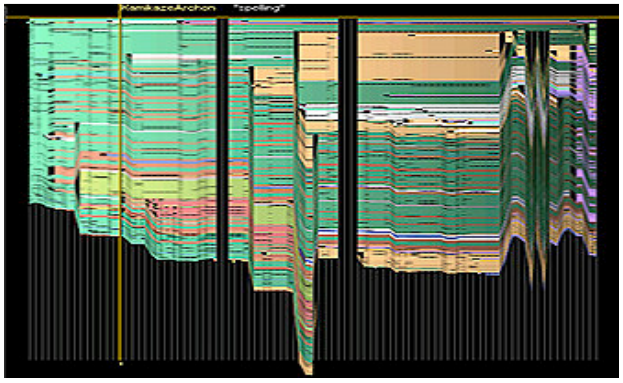

Figure1: History flow diagram showing edits made to the Abortion page until Aug. 2003.

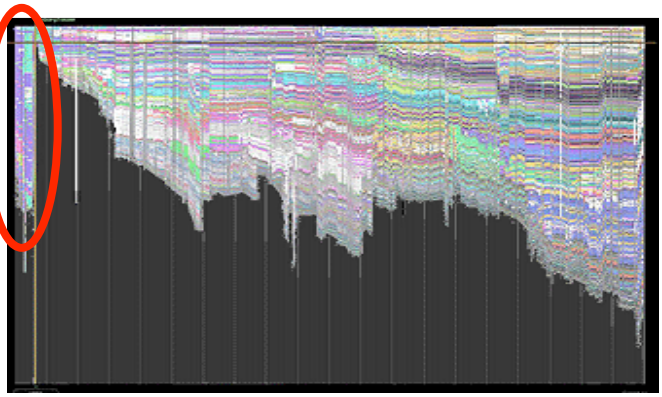

Figure 2: History flow diagram showing edits made to the Abortion page until Oct. 2005. The edits shown in the 2003 image (on the left) are highlighted by a red ellipse here.

users articulate what they perceive as the main issues of quality in the improvement of articles.

Taking a different angle, Bryant et al. focused on the social trajectories of nine active "Wikipedians," showing how their roles changed as they became more engrossed in the Wikipedia community [2]. The researchers found that, as participation became more central and frequent, Wikipedians adopted new goals, new roles, and used different tools to achieve new ends. Wikipedians usually moved from a local focus on editing individual articles to a concern for the quality of Wikipedia content as a whole, taking on more "administrative" roles in the site. One such role is that of watchdog, where users monitor community activities and look for opportunities to help and correct mistakes. Even though the study described in this paper focuses on Wikipedia as a whole—instead of focusing on individual users—some of the same trends have been identified and will be discussed in the sections that follow.

More generally, Wikipedia can be viewed as a massive experiment in collective action. There is a large literature on this topic in contexts ranging from online communities [7] to self-organizing African irrigation collectives [9]. One of the themes of such work (e.g., design principle six in [9]) is the need for a local, low-cost arenas for resolving conflicts. The hypothesis put forward in this paper is that talk pages and other coordination spaces in Wikipedia serve precisely this role.

The present study builds on the work of Viégas, Wattenberg and Dave [11]. In 2003 Viégas et al. downloaded the entire archive of Wikipedia history in order to visualize the evolution of articles and analyze conflict and collaboration patterns. Using the history flow visualization technique they identified patterns such as edit wars and vandalism repair, which were then investigated further through statistical analysis. The present study takes the same approach and tries to answer two main questions: How has Wikipedia changed since Viégas et al.'s study? What role do Talk pages play in the coordination of work?

3. Compare & Contrast

3.1 The 2005 Data

To investigate the current state of Wikipedia, we downloaded a file from the Wikipedia site that included all pages (except for deleted pages), along with full revision histories, from the October 2005 English-language Wikipedia. We refer to this full data set as FULL05. (The data from May 2003 used in [11] will be referred to as FULL03.) The data was imported into a MySQL database using tools provided by Wikipedia. Because of the large size of the database (more than 250 GB), some queries were extremely slow. To speed the process for the slowest queries, we created a 5% random sample of article pages, with full revision histories for each page in the sample. As detailed below, this slice of the data, which we refer to as SAMPLE05, was used in several of our analyses.

3.2 History Flow

As in [11], we began by applying the history flow visualization application. This tool produces a graphical view of the revision history of an individual page, plotting revision sequence on the x-axis and using the y-axis to show how the contributions of different authors are added, deleted, and rearranged over time. The application is often valuable for providing an overview of editing activity.

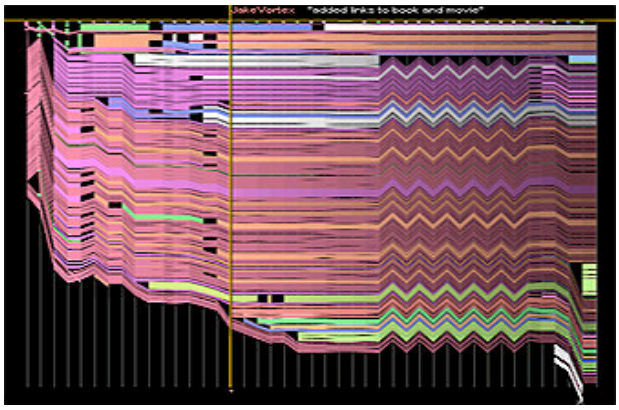

Figure 3: History flow diagram showing edits made to the Chocolate page until Aug. 2003. The presence of an edit war (the zigzag pattern on the right) is clearly visible in the diagram.

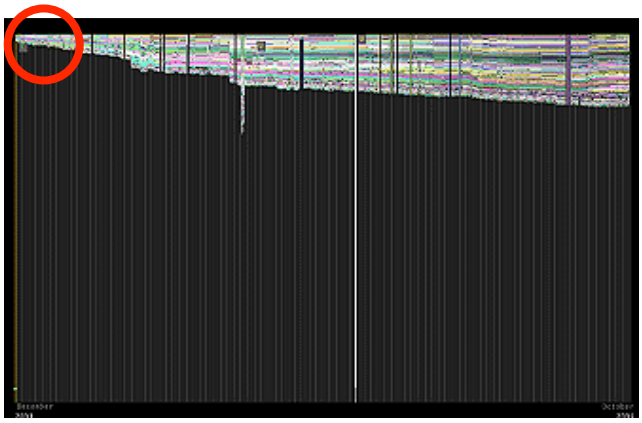

Figure 4: History flow diagram showing edits made to the Chocolate page until Oct. 2005. There have been so many edits since mid 2003 that the entire diagram shown in the previous image has become very small (circulated in red), making it impossible to distinguish the pattern of the edit war.

Adapting the tool to work with new data required several changes due to the larger size of the revision histories in FULL05. The most important was the creation of a user interface that allowed smooth zooming and panning, since many articles had more revisions than could fit on the screen. This in turn required rewriting the rendering layer of the application to use OpenGL rather than the standard Java2D package.

With these changes in place, we were able to examine many of the pages described in [11]. For example, Figure 1 shows the diagram of the edits to the page on "Abortion" from the FULL03 data set. In contrast, Figure 2 shows the history flow diagram for the same page using the FULL05 data. The area circled in red corresponds to the data in Figure 1. Similarly, Figures 3 and 4 show the history flow diagrams for "Chocolate."

Other than the dramatic change in scale, the diagrams suggested several hypotheses. As in the FULL03 data set, vandalism was evident (e.g., the spike in the Chocolate page) ; this is discussed in detail below. Similarly, page size continued to show a general upward trend with number of edits, with occasional sharp drops in size. One change was an apparent drop in frequency of "edit wars," i.e., long back-and-forth sequences of editors undoing each other's changes. One possible reason for this is the voluntary adoption within the Wikipedia community of a "three revert rule," barring each member from making more than three reverts to a given page in a 24-hour period [13].

3.3 Vandalism

One of the main results of [11] was that Wikipedia showed remarkable resilience in the face of malicious edits. Such edits (known as "vandalism" in Wikipedian terminology) were often corrected rapidly, sometimes in a matter of minutes, by members of the community.

Not only has Wikipedia grown in the years since that analysis, but it has become a high-value target. Its pages frequently are found at the top of web searches, and it is referenced in major media outlets. Given its growth and potentially higher appeal to vandals, it is natural to wonder whether the community-based repair mechanisms have kept pace.

In some situations, the old repair mechanisms have not been sufficient. For example, in 2004 the first page was protected to prevent changes [14], due to an intense level of vandalism (perhaps predictably, this page was "George W. Bush"). As of this writing that page usually remains in a "protected" mode which allows changes only by users who have been registered for a certain length of time. (In the FULL05 data, a small minority (0.09%) of pages were marked as protected.) In addition, 2006, light restrictions were placed on the creation of pages to hold vandals in check.

Despite some well-publicized problems, a statistical scan of the SAMPLE05 database shows that the basic fast-repair characteristics of Wikipedia remain strong. The results of our analysis are shown in Table 1. The median time to revert a "mass deletion" of a page was 2.9 minutes, and an "obscene" mass deletion was reverted in a median of 2 minutes. (See [11] for precise

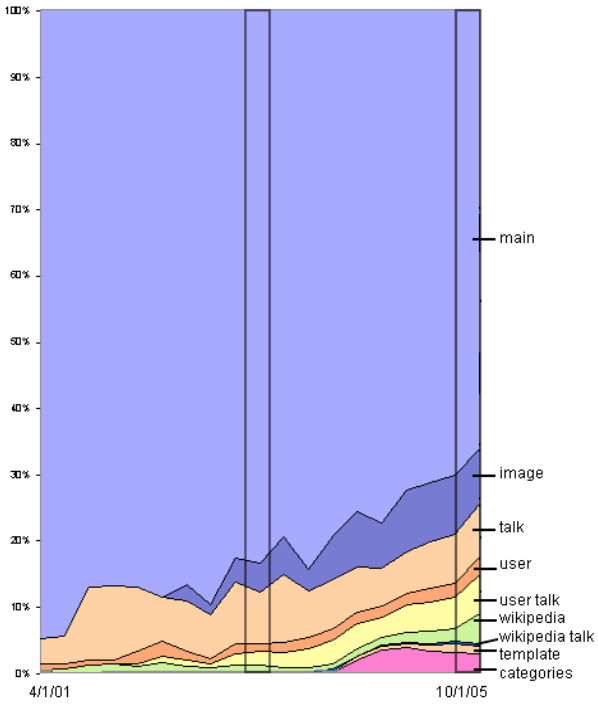

Figure 5: Evolution of Wikipedia namespaces over time, based on page count. The two selected vertical "slices" represent data in FULL03 and FULL05.

definitions of these terms.) Thus for a large set of pages, the fast-repair mechanisms continue to function. These statistics are similar to the published results for FULL03, where mass deletions were reverted in a median time of 2.8 minutes, and obscene mass deletions were reverted in a median time of 1.7 minutes. Two other comparisons with that study are worth noting. First, although the 2005 statistics were generally similar, we found a much higher median time between all edits on a given page: 726 minutes as opposed to 90 minutes. The reason for this difference is unclear. Second, we recalculated the same statistics for articles before May 2003 that still existed in SAMPLE05. Note that this subset is not the same as FULL03 since it does not include the many articles deleted since May 2003. Nonetheless, despite a built-in bias towards articles that survived for years, the recalculated statistics were generally similar to the original ones (though again the median time between all edits was much higher). See Table 2.

Table 1: 2005 Statistics (SAMPLE05)

Revision Type	Number	Mean time	Median time
All content	901,242	19.1 days	1113 minutes
Mass delete (MD)	4,848	7.0 days	2.9 minutes
MD obscene	105	0.13 days	2 minutes

Table 2: Recalculated 2003 Statistics (FULL03)

Revision Type	Number	Mean time	Median time
All content	558,702	22.4 days	726 minutes
Mass delete (MD)	3,067	11.0 days	8 minutes
MD obscene	19	0.018 days	1 minutes

4. How has Wikipedia grown?

Because the site has grown in an organic, decentralized manner (users can create and delete content at will), growth patterns can be highly telling of the ways in which the Wikipedia community is evolving. Moreover, as noted in previous studies, there are different ways in which users may contribute to Wikipedia, ranging from article writing to more administrative tasks [2]. Here we discuss statistics relating to the growth of Wikipedia, and how different sections relate to different roles.

4.1 Namespaces

When people think about Wikipedia, they tend to focus on the encyclopedia articles. It is true that the bulk of the site consists of article pages. Nevertheless, the site is much more than its encyclopedic content. In fact, Wikipedia is divided into 20 sections, called namespaces, each serving a special purpose (see table 3). Each namespace has an associated talk namespace for discussion—for instance, the namespace "Image" has "Image Talk" associated with it. For the purposes of this paper, we will focus on namespaces ranging from zero (main) to seven (image talk).

Table 3: List of all Wikipedia namespaces

#	Namespace Title	#	Namespace Title
-2	Media	**8**	MediaWiki
-1	Special	**9**	MediaWiki talk
0	(*main*)	**10**	Template
1	Talk	**11**	Template talk
2	User	**12**	Help
3	User talk	**13**	Help talk
4	Wikipedia	**14**	Category
5	Wikipedia talk	**15**	Category talk
6	Image	**100**	Portal
7	Image talk	**101**	Portal talk

The main namespace contains all encyclopedic articles, that is, the "meat" of Wikipedia. "Talk" refers to discussion pages associated with these articles. The "User" namespace provides pages for registered users' personal presentation and auxiliary pages for personal

use containing, for instance, bookmarks to favorite pages. "User talk" refers to discussion pages associated with User pages. "Wikipedia" refers to pages that explain policies and guidelines and talk about Wikipedia's Sister Projects (e.g. Wiktionary, Wikibooks, Wikinews, etc). "Wikipedia Talk" refers to discussion associated with pages in the Wikipedia namespace. Finally, "Image" is a namespace that provides information about images and sound clips, one page for each file, with a link to the image or sound clip itself. "Image Talk" is the discussion space associated with the Image namespace.

Not all namespaces have grown at the same rate over the years. In fact, some namespaces hardly existed in the beginning of the encyclopedia, having become active just recently. As can be seen in figure 5, the composition of Wikipedia has changed since the FULL03 snapshot, with non-content pages now taking up a higher proportion of the total. Table 4 shows a comparison of the FULL03 and FULL05 data, showing that between 2003 and 2005 the fastest growing namespaces were User Talk, followed by Wikipedia (guidelines). In addition to specific namespaces, coordination also happens through WikiProjects, which are collections of pages devoted to the management of a specific set of articles—e.g. the set of pages under "architecture," "music," etc. The existence of such projects plus the growth of different namespaces suggests that a growing amount of activity happens in auxiliary spaces where users coordinate action rather than edit articles. This pattern echoes the tendency of active Wikipedians to move from having a local focus—editing individual articles—to a more high-level concern for the quality of content and the health of the community, as described by Bryant et al. [2].

Table 4: Page Growth Factor per Namespace

Namespace	Pages in 2003	Pages in 2005	Growth rate
Main	170,369	1,531,095	**9x**
Talk	20,067	229,999	**11 x**
User	3,324	76,491	**23 x**
User talk	2,564	199,264[1]	**78 x**
Wikipedia	1,211	81,738	**68 x**
Wikipedia talk	441	7,267	**16 x**
Image	6,970	292,451	**42 x**

[1] In this table, the number of User Talk pages is considerably higher than the number of User pages. The reason for this discrepancy is that, when User Talk pages become too long, users archive them as separate pages.

Figure 6: Screen shot of talk page on Feminism with two separate discussion topics. The indented arrangement of postings reflects the threaded nature of the discussion. Blue links at the end of each post indicate users' signatures. Usernames have been anonymized for publication purposes.

5. Talk pages

To better understand the increase in coordination-related pages, we decided to examine one of these categories in depth. Even though coordination happens in various venues in Wikipedia—the different talk namespaces, Wikiprojects, user pages, etc—we chose to focus our inquiry on article Talk pages since they are one of the oldest coordination mechanisms on the site. Moreover, because of their direct relation to the encyclopedia entries, they seemed like a natural follow-up to [11]'s study.

In Viégas et al, Talk pages were characterized as places where conflict was resolved [11]. While it is true that they serve this function, a closer reading of the Talk pages indicates they play an important role in planning and other types of coordination as well. Editors discuss paragraphs that need reworking and sections that should be added or trimmed. In other words, editors use Talk pages as a place for collective planning as well as a platform for dispute resolution.

Before analyzing this phenomenon further, we discuss some of the basic statistics behind Talk pages. Non-empty Talk pages exist for 14.5% of the article pages in the FULL05 database. Heavily edited articles and Talk pages go hand in hand. While the average edits per page in Wikipedia is roughly 15 (median = 2), around 94% of the pages with more than 100 edits have related Talk pages. Conversely, articles with associated Talk pages have, on average, 5.8 times more edits and 4.8 times more users than articles without a

discussion forum (see Table 5). The exceptions to this association are generally pages with many edits and little discussion; many of these exceptions are simply collections of factual statistics, such a "List of Canadian Federal Electoral Districts by Region", with a 45 byte Talk page and a 12K article.

Table 5: Comparison of edits and user averages on article pages with and without talk pages in the main namespace.

	articles w/ talk page	articles w/out talk page	all articles
Avg edits	52.6	9.1	15.4
Median edits	18	2	2
Avg users	17.3	3.6	5.6
Median users	9	2	2
Count	222,150	1,308,945	1,531,095
% of total	14.5	85.5	100

5.1 Going beyond the numbers

While the numbers quoted above indicate that Talk pages play a role in the article editing process, they provide no clue as to the nature of this role. An informal examination of Talk pages indicated to us that in fact many different types of coordination take place within these pages, ranging from high-level discussion on the goals of an encyclopedia to discussions on the minutiae of etymology.

Table 6: List of talk pages that were coded for content

ID	Talk page title	ID	Talk page title
1	Baldness	**14**	Playing Card
2	Blue Whale	**15**	Polygon
3	Chess	**16**	Prussian People
4	Color Theory	**17**	R.E.M (band)
5	Continuous Function	**18**	Steganography
6	Eubacteria	**19**	Tropical Cyclone
7	Feminism	**20**	United Kingdom
8	Flag	**21**	Van Allen Radiation Belt
9	George W. Bush	**22**	Wave
10	Gmail	**23**	William Shakespeare
11	Hanging, Drawing and Quartering	**24**	X
12	Karma	**25**	Yasser Arafat
13	Online Dating		

To better understand the relative importance of these different aspects of coordination, we decided to manually classify all user posts in a purposeful sample of 25 Talk pages from the Main namespace. The sample was chosen to include a variety of controversial and non-controversial topics and span a spectrum from hard science to pop culture. To ensure the sample contained cases with thorny coordination issues, we selected some pages, such as "George W. Bush," where page protection had been necessary. The list of selected

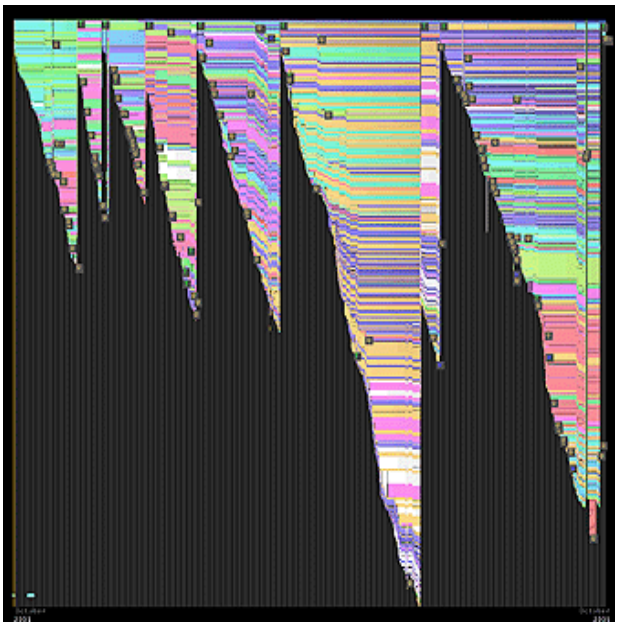

Figure 7: History flow diagram of the talk page on "God," showing the cyclical nature of archiving operations. As the increasing length of the page reaches unwieldy proportions, the discussion is archived and the live talk page returns to a more manageable size.

Talk pages and their associated code numbers can be seen on table 6. Each page was analyzed by two separate researchers to ensure coding consistency; when the classifications disagreed, they were resolved through mutual discussion.

5.1.1 Coding Methodology. Talk pages are conversation places and, as such, they are governed by different rules of etiquette from article pages—for instance, in Talk pages contributors are not supposed to erase each other's comments and every person is supposed to sign his or her postings. Despite these differences, Talk pages are produced with the same wiki technology, with no special features to support conversation threads. Instead, threading happens by social convention and users are encouraged to follow certain posting rules to ensure that conversation threads be easily recognized. The main layout and markup conventions in Talk pages can be summarized as follows:

-*Signatures*: user signatures at the end of each post are considered common courtesy. Contributors are encouraged to sign every post by linking their username to their personal user page. A special button in the editing user interface makes signatures easy.

-*Indentation*: users are encouraged to indent related postings— for instance, the answer to a question—so that the relationship between associated contributions becomes visually clear.

-*Discussion Topics*: contributors are told to place each new conversation topic at the end of the Talk page, under a different section header. This procedure makes it easier to scan the Talk page because it breaks the content into visually separate sections.

In practice, however, users do not follow these rules all the time, making it challenging to automatically identify individual postings on a given page. In the pages coded for this paper, we found that, on average, users signed their names only 67% of the time, making user signatures an unreliable method for establishing the number of individual contributions. In addition to unreliable signature patterns, Talk pages contain an enormous variety of postings: some are just a few words, others are more like essays. Some posts quote passages from the associated article while others link to outside sources and contain images to better illustrate their point. Conversation threads range from single "orphan" posts to exchanges including hundreds of contributions.

Given this wide range of conversation profiles, it became essential to find a dependable way to automatically count the amount of activity present in a given Talk page. We decided to base our calculation on the number of individual postings in a page. Unlike other measures such as the length of a Talk page, or the number of sections, this metric seemed to best capture the amount of exchange between contributors.

We relied on a combination of metrics to automatically determine the number of single postings on a given Talk page. Our algorithm defines a post as anything found after one of the following separators:

- a horizontal rule (see Figure 6)
- a signed user name
- a new indentation level

For example, the page on Figure 6 would contain seven posts by this algorithm (the end of the last post has been cut off in the screenshot).

This combination of metrics seemed to yield the most plausible results for identifying individual postings on a Talk page. It should be noted that, because postings to Talk pages are not supposed to be erased, the pages can grow indefinitely, sometimes becoming unwieldy. To remedy this problem, users are encouraged to archive the contents of Talk pages every so often. The archived pages become dormant records of past discussions, whereas the shorter, "live" Talk page continues to serve as the main conversation locus. The existence of talk archives means that, in some cases, the Talk pages coded for this study were thin slices of the discussion around a given subject. For example, the Talk page on George W. Bush had 43 archived pages associated with it. In the pages analyzed for this paper, none of the archived pages were coded.

5.1.2 Posting Dimensions. Participants' postings on Talk pages were classified along the following 11 dimensions:

1. *Requests/suggestions for editing coordination*. Postings that help users plan editing activity. E.g. "I would like to suggest pruning the external biographies list. A sampling of 5 or 6 biographical links would be sufficient." (Talk page for the article on Yasser Arafat)
2. *Requests for information*. Postings where the user requests information that is related to the article topic without a clear intention to edit the article itself. E.g. "Hey, I'm doing a report and I can't find these things: How do people vote in Great Britain/UK? What is the role of political parties? " (Talk page for the article on the UK).
3. *References to vandalism*. Postings that refer to acts of vandalism on the article page. E.g. "Could we unprotect this article for a fortnight, see if any positive anonymous and new user contributions come up? Not all anonymous users who edit this article are necessarily vandals" (Talk page for the article on George W. Bush).
4. *References to Wikipedia guidelines and policies*. References were counted whenever users pointed out official Wikipedia guidelines either by name (e.g. NPOV[2]) or by linking to policy pages. E.g. "I am concerned that this article is too harsh and accusatory. I will be making some edits to aim at a more genuine NPOV tone where possible" (Talk page for the article on George W. Bush).
5. *References to internal Wikipedia resources*. Postings that link to other Talk pages or archives to illustrate/explain a user's comment. E.g. "Cleanup done. See User:Samboy/Chess_zapped [linked] to see what information I removed from the article when cleaning it up" (Talk page for the article on Chess).
6. *Off-topic remarks*. Postings unrelated to the article. E.g. "I currently have 6 Gmail invites and I want to give 3 away to Wikipedia members" (Talk page for the article on Gmail).
7. *Polls*. Voting sessions organized by users to decide on controversial editing actions. E.g. "The vote is this: Should the above paragraph be included in

[2] NPOV = "neutral point of view," Wikipedia's main editing policy.

the article? The four possible answers are: Yes, No, Abstain, and Other" (Talk page for the article on Chess).

8. *Requests for peer review*. Users hoping to elevate articles to "featured" status may solicit a peer review.
9. *Information boxes*. These are call-out boxes placed at the top of the Talk page indicating any special processes the article page may have gone through—for instance, if the page was a featured article at any point, or whether it is currently being peer reviewed—or whether the page is part of a given WikiProject.
10. *Images*. Image files posted on the Talk page.
11. *Other*. Postings that do not fit any of the above categories. Not all of the above dimensions are mutually exclusive; a single post may be marked as being relevant to more than one dimension. For instance, a request for coordination may point to an official Wikipedia guideline page. Such a post would be marked both as "request for coordination" and as a "reference to Wikipedia guidelines and policy." Likewise, if a request for editing a paragraph in the article page is accompanied by an image that the user is hoping to use to illustrate the paragraph in question, the post would be marked both as "request for coordination" and as containing an "image."

5.2 Results

The selected Talk pages ranged in size from 12K to 128K. The number of posts on each page ranged from 5 in Color Theory to 205 in Gmail.

Requests for coordination were, by far, the most common kind of posting, accounting for over half of the contributions on Talk pages (see table 7). This establishes the crucial strategic role that Talk pages serve in Wikipedia. Contributors use Talk pages to discuss their editing activities in advance, to ask for help, and to explain the reasons why they think specific changes should be made.

Next in frequency were requests for information, which occurred once in every ten posts. Such requests are usually made by visitors who have no intention of editing the associated article and they suggest the use of Talk pages as a place to tap into expert knowledge of specific topics. Users who post such requests for information seem to perceive contributors to Talk pages as an approachable community of experts. A frequent response from Talk page "experts," however, is to direct such users to pages or archived discussions in Wikipedia that might answer their questions. In all 25 pages coded for this paper, the overwhelming majority of requests for information were answered, strengthening the sense of a strong, supportive community.

Perhaps surprisingly, most Talk pages contain relatively few off-topic remarks. The three pages with the highest proportion of irrelevant postings were Online Dating, Gmail, and Steganography. Interestingly, all three are related to current technologies and most of the off-topic postings were messages where users shared interesting bits of trivia or new experiences they had just had with the technology. For instance, in the Steganography[3] Talk page, a participant posted a puzzle: he placed and image with a hidden message and asked others to decipher it. In the Online Dating Talk page, users posted remarks about their personal experiences with different online dating systems, and one user described how she found her "life mate" online.

The next notable category of postings is made up of references to official Wikipedia guidelines, which account for 7.9% of the activity in Talk pages. This indicates that policies and guidelines are actively used by the Wikipedia community. In fact, a pattern became clear: in Talk pages containing serious disagreements or flame wars, moderators would often refer to official policies. Consider the following passages from the Talk pages on "Feminism" and on "R.E.M. (band)" respectively:

Please read Wikipedia:No Personal Attacks [link goes to guideline page]; note that "Racial, sexual, homophobic, religious or ethnic epithets directed against another contributor" are included in the category of "personal attacks." Also note this section: "Using someone's affiliations as a means of dismissing or discrediting their views—regardless of whether their affiliations are mainstream or extreme"

The real problem: Wiki articles are not supposed to include what's referred to as Wikipedia:Original Research [link goes to guideline page]. Basically, if a sourced article didn't make the claim, we can't include it. Example: this article is the first time I've heard early R.E.M. compared to the Who. Regardless of my opinion on the validity of such a claim, if someone else didn't make the claim in a citable article, we can't include it here. (Read through the"original research" link above and see what I mean)

In other words, when discussions became uncivil, references to the relevant guidelines and policies were made, perhaps with the intention of bringing the conversation back to a more courteous level. (Generally, this strategy seemed to work.) Often

[3] Steganography is the art of including hidden messages in text or images in such a way that no one apart from the intended recipient knows of the existence of the message.

Wikipedians would explain the guidelines, as in the R.E. M quote above, so that the references seemed to serve an educational function. Finally, pointers to other internal resources such as associated Talk pages or archives occurred in 5.4% of all postings. Unlike references to guidelines—coded separately, as seen above—these links seemed not to be aimed at getting contributors to behave properly but rather at pointing to key passages from other pages that are relevant to the current conversation. This means that users find it important to refer to existing content instead of paraphrasing passages or taking quotes out of context, as can be seen in this passage from the Talk page on William Shakespeare:

As an experiment, I'm going to semi-protect this page for a bit so only logged-in users can edit it. This appears to be done more on Wikipedia of late due to increasing vandalism. (However, opinion is very split over this, just see here [Wikipedia_talk:Semi-protection] and here [Wikipedia: Village pump (policy)]). If anyone has a problem with this please let me know.

Table 7: Distribution of postings on talk pages

	Average	Maximum value *(page title)*	Minimum value *(page ID)*
Coord. req.	**58.8%**	**97%** *Yasser Arafat*	**2%** *Van Allen rad. belt*
Info req.	**10.2%**	**57%** Flag	**0%** *4, 6, 13, 17, 21*
Off-topic remarks	**8.5%**	**60%** *Online Dating*	**0%** *1, 2, 3, 4, 6, 8, 14, 16,*
Ref. to guidelines	**7.9%**	**39%** *R.E.M, Feminism*	**0%** *4, 5, 8, 13, 18, 22, 24*
Internal resources	**5.4%**	**17%** *William Shakespeare*	**0%** *1, 4, 6, 8, 14, 15, 16,*
Ref. to vandalism	**3.5%**	**36%** *Prussian People*	**0%** *1, 2, 4, 5, 7, 8, 10, 11, 13, 17, 18, 20, 21*
Info box	**1.6%**	**6%** *Blue Whale*	**0%** *4,5,10, 11, 12, 13, 15, 16, 17, 22, 24*
Polls	**0.4%**	**9%** *Chess, United Kingdom*	**0%** *all other pages*
Peer review req.	**0.3%**	**3%** *Flag, Wave*	**0%** *all other pages*
Images	**0.2%**	**2%** *Blue Whale*	**0%** *(all other pages except for 10 and 19)*

6. Discussion

Wikipedia is becoming a more diverse and complex site, with different kinds of pages serving distinct purposes. With a greater proportion of the site being devoted to "meta" pages, more users seem to be involved in organized editing activities. Talk pages constitute a key place of editing coordination. With the creation of a Talk space for every namespace in Wikipedia, the community has ensured that there can be behind-the-scenes coordination in every aspect of the site.

It has been hypothesized that one reason why Wikipedia works well is its "neutral point of view" policy (NPOV) [3][10][11]. Having a policy, however, does not ensure that members of a community will abide by the rules. An important step for any community is to create mechanisms—both formal and informal— through which policies can be taught and enforced. In this paper, we have described how references to official policies on Talk pages constitute just this sort of informal mechanism. We found several cases in which exchanges in Talk pages verged on arguments or personal attacks, at which point "moderators" took a step back and reviewed the ways in which article edits ought to happen in Wikipedia. These users thus turned the spotlight from the offending parties onto Wikipedia policies. Because Wikipedia has dozens of specific guidelines—ranging from image use policy, to policies on harassment, to "Wikipedia is not a publisher of original thought" [15]—it is easy for moderators to point users to the precise rules they might be breaking. In the pages coded for this study we came across some heated exchanges but there was never a case that could not be controlled by the community and its collective resolve in pointing contributors to official Wikipedia policy pages. It is notable that the George W. Bush page may need protection, yet its Talk page does not.

We also observed many instances in which guidelines were linked to from Talk pages after new users breached rules without being aware of them. By pointing out how things should be done and linking to official policy sources, experienced users help educate newcomers about the culture of Wikipedia and what is expected of contributors.

7. Conclusion and future directions

We have presented the results of an empirical study of Wikipedia. Our results show how Wikipedia has grown and evolved since an earlier study based on 2003 data. While the scale of the encyclopedia has

increased dramatically, several of the findings from [11]'s investigation remain correct. Certain types of vandalism, for example, continue to be fixed within minutes. On the other hand, some aspects of the encyclopedia have changed. One significant difference is the proportion of pages devoted to coordination and administration.

To better understand how these pages function, we focused our attention specifically on the Talk pages: places for discussion that are attached to each article in the encyclopedia. Starting from a process of manual classification and coding of a sample set of Talk pages, we found that a significant amount of planning occurs on these pages. This contrasts with the analysis in Viégas et al, which presented Talk pages as an arena for retroactive resolution of disputes.

Furthermore, we found that conversation on Talk pages is in some respects formalized and policy-driven. Special etiquette has evolved for the Talk pages, and explicit references to written guidelines are frequently invoked. Overall, the kind of process and policy enforcement that happens in Talk pages seems to play a crucial role in fostering civil behavior and community ties.

Our results point to several directions for further investigation. The richness of the conversation around Talk pages suggests that additional study would be fruitful. It would be especially helpful to find ways to automate or partially automate the analysis of Talk page content: by applying text analytics, for example, or creating new visualizations that would link the discussion with the article it discusses.

It would also be natural to investigate other "meta" sections of Wikipedia. The User and User Talk namespaces have grown tremendously, and hold much of the community interaction. One of the mysteries of Wikipedia is the motivations of the volunteers who donate their time; it has been hypothesized [4] that peer recognition plays an important role. An examination of the User and User Talk areas might shed light on this conjecture.

Finally, it would be worthwhile to investigate the evolution of the policies and processes that serve as reference points during discussion. Such trappings of bureaucracy are often seen as the result of the exertion of power from the top down, yet in Wikipedia they seem to emerge, to some degree, spontaneously. It would be fascinating to explore whether Wiki technology— seemingly antithetical to bureaucracy— actually supports or even encourages conventional forms of organization.

8. Acknowledgments

We are grateful to the members of the Collaborative User Experience group at IBM for helpful comments on this work.

9. References

1. Benkler, Y. Coase's Penguin, or, Linux and The Nature of the Firm. The Yale Law Journal, Vol 12, N 3. December 2002.
2. Bryant, S., Forte, A., Bruckman, A. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Enciclopedia. In Proceedings of GROUP 2005.
3. Emigh, W., and Herring, S. Collaborative authoring on the Web: A genre analysis of online encyclopedias. In Proceedings of HICSS-38, 2005.
4. Forte, A., Bruckman, A. Why do People Write for Wikipedia? Incentives to Contribute to Open-Content Publishing. GROUP 05 workshop position paper.
5. Giles, J. Internet encyclopaedias go head to head. In Nature, 14 December 2005. http://www.nature.com/news/2005/051212/full/438900a.html
6. Holloway, T., Bozicevic, M., Börner, K. Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors. Submitted to Complexity, 2005.
7. Kollock, P., and Smith, M. Managing the Virtual Commons: Cooperation and Conflict in Computer Communities. in Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives, Susan Herring (ed), Pp. 109-128. Amsterdam: John Benjamins. 1996.
8. Lih, A. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for Evaluating Collaborative Media as a News Source. Proceedings of the Fifth International Symposium on Online Journalism, 2004.
9. Ostrom, E. Collective Action and the Evolution of Social Norms. Journal of Economic Perspectives, Vol. 14, No. 3, pp. 137-158, 2000.
10. Stvilia, B., Twidale, M., Gasser, L., Smith, L. Information Quality Discussions in Wikipedia. Technical Report ISRN UIUCLIS--2005/2+CSCW, 2005.
11. Viégas, F., Wattenberg, M., & Dave, K. Studying Cooperation and Conflict between Authors with history flow Visualizations. In Proceedings of SIGCHI 2004.
12. Voss, J. Measuring Wikipedia. In Proceedings 10th International Conference of the International Society for Scientometrics and Informetrics, 2005.
13. http://en.wikipedia.org/wiki/Three_revert_rule
14. http://en.wikipedia.org/wiki/Wikipedia:Protected_page
15. http://en.wikipedia.org/wiki/Wikipedia:List_of_policies