A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia

Aaron Halfaker
Grouplens Research
University of Minnesota
200 Union St. S.E.
Minneapolis, MN 55455
halfak@cs.umn.edu

Aniket Kittur Robert Kraut
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
{nkittur, robert.kraut}@cs.cmu.edu

John Riedl
Grouplens Research
University of Minnesota
200 Union St. S.E.
Minneapolis, MN 55455
riedl@cs.umn.edu

ABSTRACT

Wikipedia is a highly successful example of what mass collaboration in an informal peer review system can accomplish. In this paper, we examine the role that the quality of the contributions, the experience of the contributors and the ownership of the content play in the decisions over which contributions become part of Wikipedia and which ones are rejected by the community. We introduce and justify a versatile metric for automatically measuring the quality of a contribution. We find little evidence that experience helps contributors avoid rejection. In fact, as they gain experience, contributors are even more likely to have their work rejected. We also find strong evidence of ownership behaviors in practice despite the fact that ownership of content is discouraged within Wikipedia.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems

Keywords

Wikipedia, Peer, Peer Review, WikiWork, Experience, Ownership, Quality

1. INTRODUCTION

Since its founding in 2001, Wikipedia has become one of the most ubiquitous sites on the Internet. Web searches on almost any topic yield at least one reference to Wikipedia in the first page of results. Perhaps because of this prominence in search results, wikipedia.org is one of the top ten most visited domains on the Internet by many measures. Most remarkably, nearly all of the content in Wikipedia is contributed by volunteers. The functioning of this group of millions of volunteers is noteworthy, in that they work with very little formal organization. In this research, we seek to understand how the quality of content is modulated by the experience of these teams of volunteers and by their feelings of ownership.

One of the key components of Wikipedia is the review process through which contributions are rejected or accepted. This process is informal and, to an outsider, appears disorganized, with its reliance on watchlists and Internet Relay Chat channels. However, the review process is robust and effective in practice: 42% of vandalistic contributions are repaired within one view and 70% within ten views [15].

Many other systems use peer review, though usually in a more structured manner. For instance, conferences typically have three peers of the authors read each submitted article to decide whether it should be accepted or rejected. Similar peer review systems include NSF grant panels and arts competitions. The goal of these review processes is to ensure that high quality work survives while lower quality work is rejected. One important research question is how effective these peer review processes select for high quality contributions.

Wikipedia is generally not thought of as a peer review system since any contribution can be made and saved instantly, but Stvilia et al. explained that the open editing system constitutes an informal peer review that moderates the quality of articles [16]. In this research, we explore the effectiveness of the peer review system within Wikipedia by examining how the characteristics of editors and their changes predict which contributions will be rejected.

Previous work in evaluation of formal peer review systems either determined the quality of reviews by exper evaluation [9] or checking for significant differences in reviewer evaluations [18]. In either case, the evaluation focuses on a system's ability to select for high quality content despite bias. In this research, we seek to determine whether the results of Wikipedia's peer review process are primarily driven by the quality of work or whether non quality-related factors are also influential.

There are three contributions of this paper to the state of the art. First, we develop an automated measure (word persistence) for evaluating the quality of individual contributions. Using this automated measure, we examine whether words that have become established as high quality are difficult to change. We also test whether the recent quality of editors' work predicts whether their new work will be rejected.

Second, we look at the experience of an editor as a predictor of whether the contribution will be rejected. Decades of research show that individuals, groups and organizations

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

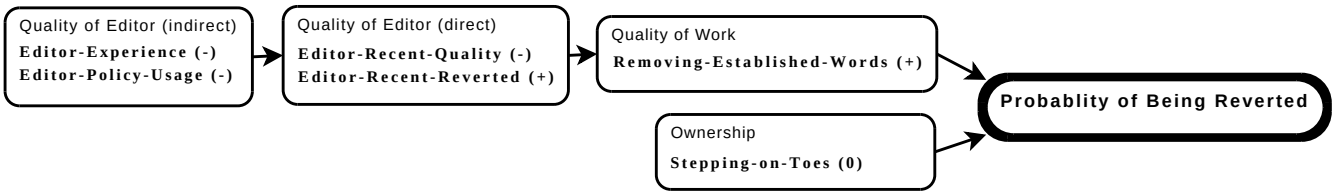

Figure 1: How various factors would effect the probability of being reverted in an ideal system. (+) represents a positive correlation, (-) represents a negative correlation and (0) represents no correlation.

all exhibit "learning by doing", where by their ability to perform complex tasks improves with their experience (see [3] for a review). We explore whether Wikipedia editors exhibit such a learning effect and quantitatively refute the premise that contributors produce more acceptable work as they gain experience.

Third, we show evidence that ownership has a strong, independent effect in Wikipedia. We have discovered that the number of toes that are stepped on by a contribution — i.e., the number of editors who would be likely to notice that an edit has removed a word which they had added — is a powerful predictor of whether that contribition will be rejected independent of the quality and experience of the editor making the contribution. Since the ownership of content is openly discouraged [22], this result demonstrates a non quality-related factor which has a strong effect on the outcome of review.

The concepts of ownership, experience and quality are rich ideas and no one paper can effectively explore all of them. We only explore them within the context of Wikipedia. Since many systems have similar policies and review mechanisms to Wikipedia, we believe that these findings are relevant to many other systems but do not demonstrate that relevance within this paper.

The rest of this paper is organized as follows. First, we provide our set of hypotheses in the context of related work. In the experimental methods section, we describe how measures of quality, experience, ownership and active editors are used in our analysis. In the following section we present results and discussion for each of the six hypotheses and summarize the explanatory variables through a logistic regression model. Finally, we close with conclusions and future work.

2. HYPOTHESES

To frame the hypotheses, let's define a few terms. An **edit** is the act of making and saving changes to an article. A **revision** is a state in the history of an article — i.e., edits are transitions between revisions. A **revert** is a special kind of edit that restores the content of an article to a previous revision by removing the effects of intervening edits.

We test six hypotheses that examine two categories of factors that could predict which revisions will be reverted: (1) measures of quality (direct or indirect) and (2) factors unrelated to quality. Figure 1 is an illustration of how a wiki review system would work in an ideal world. In this model, key attributes are predicted to increase or decrease the probability of a revision being reverted. Direct and indirect measures of the quality of work should effect changes to the probability of a revision being reverted while factors that do not measure the quality of work should not have a significant effect. The first hypothesis predicts that edits that remove high quality existing work are more likely to be reverted, because they are likely to reduce the quality of the article. The next two hypotheses predict that editors who have recently performed high quality edits are less likely to be reverted (direct measures of editor quality). The following two hypotheses predict that editors who have relevant experience are less likely to be reverted (indirect measures of editor quality). The last hypothesis predicts that ownership of removed words, which in the preferred model of Figure 1 should have no effect, has an effect on the probability that a revision will be reverted.

2.1 Quality of Content Changed

The purpose of peer review systems is to select for high quality work. If editors of Wikipedia select for high quality content and against low quality content in general, words that survive many subsequent revisions should be part of a higher quality contribution than words that last fewer revisions. Thus, edits that remove words that have become established are likely to be reverted since they would be removing high quality content.

HYP *Removing Established Words*: Edits that remove established words are more likely to be reverted.

2.2 Direct Editor Quality

Previous research has explored what work is most valued by Wikipedia editors [4]. Other research has found that the structure of editor contributions affects the perceived quality of articles [10]. Priedhorsky et al. built a metric for accessing the value of an editor's contributions in terms of the number of times a word is viewed [15]. However, there is very little research that has directly explored the quality of an editor's individual contributions. In this research, we build on past measures of quality, value and reputation in developing our own automated metric for the quality of a contribution in order to determine if an editor's recent record of quality predicts whether a new revision will be reverted. Previous work suggests that quality is not always a predictor of peer acceptance. For instance, Cole et al. found that the previous funding rate of an NSF applicant was not highly correlated with the probability of the current application being funded [8]. We test whether a similar property holds true for Wikipedia.

HYP *Editor Recent Quality*: Editors with a history of high quality contributions are less likely to be reverted.

Since Wikipedia is open for anyone to edit, and because articles tend to attract people with different viewpoints, conflict between editors is a common-place phenomenon. In order to provide a better understanding of the opposing groups in a conflict, Kittur et al. [11] and Brandes et al. [5] devel-

oped visualization techniques designed to render the "sides" of content-related conflict. Kittur et al. went on to suggest that conflict is not always a purely negative activity in peer collaboration systems and provides an analysis of the cost of coordination within Wikipedia.

Vuong et al. differentiates between the conflict that occurs between users and the controversiality of disputed articles by developing models that account for the aggressiveness of editors to determine how their actons should be interpreted [21]. They develop and compare various computational models that could be used to detect the difference between conflict over content and conflict due to editor personality. If peristent properties of editors, such as knowledge, skill or personality, are related to the quality of an editor's work, then we should see that the probability of a revert is a property of an editor and that an editor's recent history should be a good indicator of this property.

HYP *Editor Recent Reverted*: Editors who have been reverted recently are likely to continue to be reverted.

2.3 Indirect Editor Quality

Individuals gradually build up expertise over time, not only increasing in the complexity and amount of knowledge accumulated but also developing qualitatively different ways of organizing and representing knowledge that increases their performance [7]. Domains as diverse as automotive manufacture, pizza delivery and medicine all demonstrate a "learning effect", in which practitioners get better with experience. For example, individual surgeons, small surgical teams and large hospitals all get better at performing particular types of surgery, with higher success rates and fewer complications, the more they perform them [2]. While most prior research shows learning effects such as these, Cole et al. found that in National Science Foundation peer review decisions, an applicant's number of years of experience does not strongly predict probability of receiving funding [8]. If Wikipedia editors do become more effective editors as they gain experience, we should see a learning effect in Wikipedia.

HYP *Editor Experience*: Editors with more experience are less likely to be reverted.

Over the past few years, the way editors interact in Wikipedia and exert control over the actions of other editors has recieved a lot of attention. In a study performed over Wikipedians, those editors who become expert maintainers of the Wikipedia, Bryant et al. interviewed several of the most prolific editors to examine their motivations and growth [6]. Kriplean et al. [13] and Beschastnikh et al. [4] explain and quantitatively present the use of policy and other internal mechanisms by editors to encourage, explain and discourge various community behavior. Similarly, in an analysis of talk page[1] activity, Viégas et al. found that the majority of Wikipedia's recent growth has taken place in the coordination mechanisms and that the majority of talk page activity is dominated by requests for coordination. They conclude that these are the reasons that the system continues to maintain its strong emphasis on "coordination, policy and process" in the face of extreme growth and popularity [20]. We explore the importance of an editor's command of Wikipedia policy.

[1]Every article in Wikipedia has a talk page which is intended to be used for communications between editors about the article and editors' work.

HYP *Editor Policy Knowledge*: Editors who cite policy often are less likely to be reverted.

2.4 Ownership

Some peer contribution systems use ownership for decision-making. For example, some open source projects use implicit ownership to fill the roles of primary decision makers. In a case study of Apache software projects, Mockus et al. found that developers who had created or maintained a specific portion of code extensively were given greater say in what changes would be made to it [14]. Although ownership of content is openly discouraged in Wikipedia [22], Kriplean et al. and Thom-Santelli et al. found that there are editors who assert ownership over articles and use their previous work on an article to exert control over which contributions will be accepted [13, 17]. If the ownership of removed words has an independent effect on the probability of being reverted, that would be evidence of an inconsistency with Wikipedia's policies.

HYP *Stepping on Toes*: Edits that remove the words of active editors are more likely to be reverted.

3. EXPERIMENTAL METHODS

For our analysis, we used a random sample of approximately 1.4 million revisions attributed to registered editors (with bots[2] removed) as extracted from the January, 2008 database snapshot of the English version of Wikipedia made available by the Wikimedia Foundation[3]. In the results, we compare our analysis of the entire sample with various interesting subsamples such as those containing only non-vandalism related revisions or those containing only revisions attributed to experienced editors. To determine the independence and effect of the 12 variables analyzed, we combine them into a logistic regression with a boolean outcome variable representing whether a revision was eventually reverted[4]. Where we plot a probability of being reverted, we include a 95% binomial proportion confidence interval.

3.1 Estimating the quality of a contribution

Quality of a word. The quality of work in Wikipedia is difficult to measure. The closest metric to a gold standard for article quality is the Wikipedia 1.0 Assessment rating, an evaluation of the quality of an article which is usually attached by Wikipedia project groups interested in the article. However, as of November 2007, only approximately 25% of articles in Wikipedia were assessed a rating[5] and only 5% of articles had a rating higher than "start" [10], a rating for "mostly incomplete" articles. Even if the assessments were more pervasive, they are rarely updated and do not suggest which editors contributed positively to a change. Barnstars are a community stamp of approval that is awarded to an editor by other editors. Kriplean et al. used the attribution of Barnstars among users to discover what types of work were most valued by other editors [12]. However, Barnstars

[2]A bot editor is a computer program that performs maintenance on the pages of Wikipedia. A bot's actions are not directly controlled by a user, so we exclude them from our analysis.

[3]Database snapshots are made publicly available at http://download.wikimedia.org/enwiki/

[4]Self-reverts, where an editor reverts himself, were not counted as reverts.

[5]http://en.wikipedia.org/w/index.php?oldid=255031288

suffer from similar problems in that they are rarely given and often do not suggest which individual edits are being praised.

For our analysis, we need an automated mechanism that can be applied to a sequence of edits by an editor in order to estimate the quality of work that editor has recently produced. This metric must be available for any contribution to any article.

In forming such a metric, we make the assumption that a good estimate of the quality of a contribution to Wikipedia is the lifespan of its words. Past research has made use of several different measures of the lifespan of a word. Adler and Alfaro measured the number of seconds a word persists [1]. Priedhorsky et al. estimated the number of views of the article with a word in it [15]. We use a different metric: the number of editors who changed the article without removing the word. We prefer the number of revisions over the number of seconds because low quality words may survive many months without careful consideration in articles that are seldom revised. We prefer the number of revisions over the number of views because each revision comprises a *critical review* of an article by an active editor. Our underlying assumption is that the more reviews a contribution survives, the higher its quality. Therefore, our measure of lifespan is the number of revisions that a word survives.

To study a contribution across time, we study the lifespan of the individual words added by a contributor. The Persistent Word Revisions (PWR) of a word is the number of revisions the word persists before it is removed. In order to compute the PWR metric, we must first define what will be considered a word in a Wikipedia article. Previous work by Priedhorsky et al. and Adler and Alfaro limited the words that would be measured to only non-stopwords that occur in the article. We also include an article's wiki markup code[6] in the words we would consider since, like words that are directly rendered into the article, wiki markup is added and removed in the same way as normal content and can, therfore, can be reviewed in the same way.

Since a revert restores the state of an article, our algorithm keeps a history of words and their attribution in order to be able to reactivate words (so they may continue accruing revisions) if they are part of a revision that is reverted back to. Reverts can take two general forms: *identity reverts*, where the text of a revision is identical to a previous revision and *effective reverts*, where the effects of a previous edit are removed, but the resulting text does not exactly match that of any previous revision. For this research, only identity reverts are used due to two key advantages: comparing the raw text of a revision to previous revisions is computationally simple and determining exactly which editors' revisions were lost due to the revert is straightforward. Previous work suggests that detecting reverts in this manner includes 94% of all actual revert activity [11].

Our mechanisms for finding difference between revisions, the attribution of words to editors that add them and reattributing words when reverts occur matches the methods used by Priedhorsky et al[15]. The key difference between their measure of the persistent word *view* and our measure of persistent word *revision* is that, rather than measuring the views that take place during the life of a word, we count the revisions in which the word continues to persist.

[6]A Turing complete language used in the MediaWiki software for performing computations during page loads. Wiki markup is often used to add templates, tables and other functionality to Wikipedia articles.

Quality of a sequence of edits. As a measure of the average quality of an entire edit we use the average of the PWR over the words in the edit (PWR per Word, or PWRpW). Likewise, as a measure of the average quality of a sequence of edits, we use the average of the PWR over the words added by those edits. This average over a sequence of words, as opposed to a sequence of edits, enables us to scale the results for the number of words added during an edit. For example, an edit in which 100 words are added will have more of an effect on the average quality of contributions than an edit that adds only ten words. Equation 1 describes our approach to computing PWRpW. For simplicity, the rest of this paper will refer to PWRpW as *word persistence*.

This metric is, of course, not perfect: the meaning behind the review of a contribution depends on the state of the article and the expertise of the editor acting as a reviewer. This calculation assumes that all words in an article have the same probability of being reviewed during an edit. Words closer to the beginning of an article might be reviewed more often than words towards the end. Further, words added earlier in an article's life will have the opportunity for more reviews than words added later. In order to lessen the effect of the former assumption, we determine the quality of an editor's work over a sequence of edits to average over many different word locations. To test the validity of this assumption, we controlled for the number of revisions left in an article by subsampling based on the amount of reviews possible after a current revision. We found no appreciable difference between the usefullness of PWR in our subsample and simply taking the log of the PWR across the full sample.

$$\text{PWRpW} = \frac{\sum_{word}^{\text{words}} \log \text{PWR}(word)}{|\text{words}|} \quad (1)$$

Word persistence: ***The average number of revisions that a group of words survives.***

Verification. To check our assumption that word persistence is an appropriate measure of quality, we performed an analysis to determine if the quality of the articles edited by higher word persistence editors would be more likely to increase in their Wikipedia 1.0 Assessment than those edited by low word persistence editors. We performed a regression that mimicked the one performed by Kittur et al [10], with the addition of the scaled average persistent word metric. We found that a rise of one standard deviation average word persistence across editors active during a six month time period of an article predicted a $1/10^{th}$ assessment grade rise independent from the structure of editors contributions, the number of words added and all other predictors tested. Although this effect may appear small, it is important to note that 90% of samples showed no increase in assessment grade during the six months observed. This result supports our assumption that word persistence measures the quality an editor's contributions.

3.2 Measuring experience

There are several ways in which previous experience can be measured within Wikipedia since the database snapshot makes all editors' actions within the system available for

study. We are interested in three characteristics of an editor's history: the amount of time that an editor has been using the system (tenure), the number of interactions an editor has had with the system (previous sessions) and the number of times an editor cites policy while communicating with other editors.

We measure the number of interactions an editor has had in the system by grouping edits together into sessions. We define a session as a sequence of edits by an editor on a single page that take place in the time span of less than an hour[7]. We collapse edits in this way to control for editors that make several intermediate saves while performing one general change to an article.

In order to capture citations to policy that an editor has made, we scanned a history of all words added to talk pages and all comments attached to article edits. Since the number of citations to policy in talk pages is highly correlated to the number of policy citations in edit comments, we use only talk page citations in our model.

3.3 Measuring ownership and active editors

In order to test **HYP** ***Stepping on Toes***, we needed to determine how many active editors have their words removed by an edit. We required two mechanisms: a way to associate a word with its original creator and a way to determine which editors are active in an article at any given time. As mentioned in section 3.1, we mimic the approach used by Priedhorsky et al. [15] to attribute word reviews to an editor. This allows us to associate editors with the individual words which they have added to an article and is actually part of the word persistence computation.

Determining the an editor's status as "active" in an article was less straightforward. Since the watchlists of editors are not included in the database snapshot provided by Wikimedia, we consider an editor as active in an article if that editor has made an edit to the article or its associated talk page within the previous two weeks. This measure is advantageous over using the watchlist in that it requires active editors to be *actively* visiting and contributing to Wikipedia. For example, when editors stops editing Wikipedia, articles continue to remain on their watchlists unless they manually return to removed them. By using recent activity to determine which editors are watching, it would be impossible for us to assume that a user is active if they have not viewed the article recently.

4. RESULTS AND DISCUSSION

4.1 Quality of work

HYP ***Removing Established Words.*** In order to measure how established a word has become, we use the number of revisions that occur to its article without removing it—i.e., how long the word has persisted despite other changes to the article. This measure represents the peristence of a word at the time of its removal. In order to determine how established a set of removed words had become, we used the word persistence algorithm described in Section 3.1. Since this measure should be independent of the number of words changed during an edit, we included the number of words added and removed in our model (summarized in Section 4.5) to control for large amounts of change.

[7]An hour was chosen due to the observation that many online systems use an hour as a timeout period to reset authentication sessions.

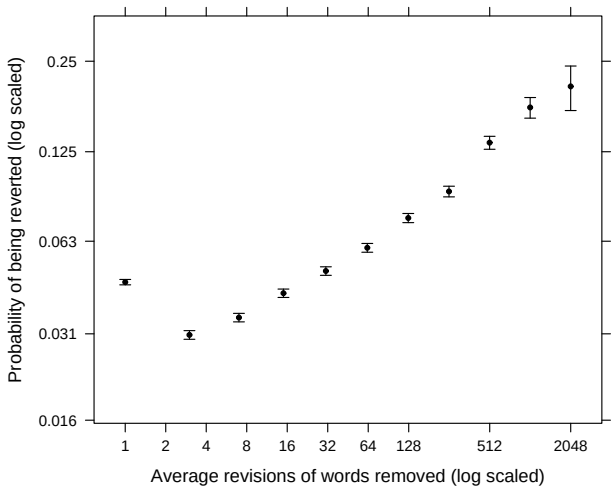

Figure 2: The probability of being reverted by the average persistence of words removed (as measured by PWRpW)

We discovered a significant increase in the probability of being reverted as the average persistence of removed words increases independent of the number of words changed. Figure 2 shows a trend where the probability of being reverted increases logarithmically as the average number of revisions that the removed words have survived increases logarithmically. There is an initial spike in the probability of being reverted for removing very young words. This uncharacteristic data point suggests that an edit that removes the words which were only just added by a previous edit is exceptionally likely to be reverted. This phenomenon could be explained by editors reacting negatively to the immediate removal of the words which they had just added.

This result supports **HYP** ***Removing Established Words*** and also provides further evidence that the word persistence metric is actually measuring the quality of a contribution — that the more revisions a word survives, the higher quality a contribution it is a part of. So long as editors value higher quality content over lower quality, the longer a word persists, the less likely other editors are to accept its removal.

4.2 Direct editor quality

HYP ***Editor Recent Quality.*** As our measure of the quality of a contribution, we use the word persistence metric described in Section 3.1, that averages the persistence of the words over a span of contributions. There are several attractive measures for defining what contributions will be considered "recent". The most direct approach is to measure the persistence of words added by the editor over a fixed timespan, such as the last week in an editor's life. Unfortunately, this measure proves a poor predictor of subsequent reverts. We speculate that the reason this measure fails is that it does not measure a constant unit of activity. One editor may have edited hundreds of articles in the past week, while another editor had not visited Wikipedia at all. Therefore, we normalized the measure by using the average persistence of words over a fixed number of edits, rather than over a fixed time period. In a sense, this metric is separating the flow of an editor's Wikipedia-time from the

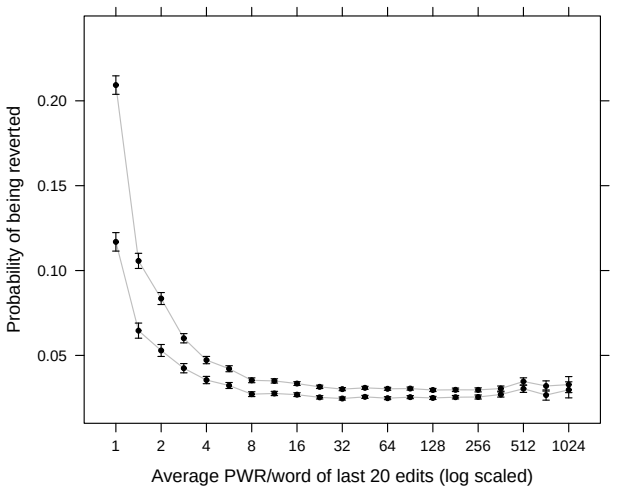

Figure 3: Probability of being reverted by the average persistence of words added in the last 20 edits. The top points represent the full sample while the bottom are controlled for vandalism.

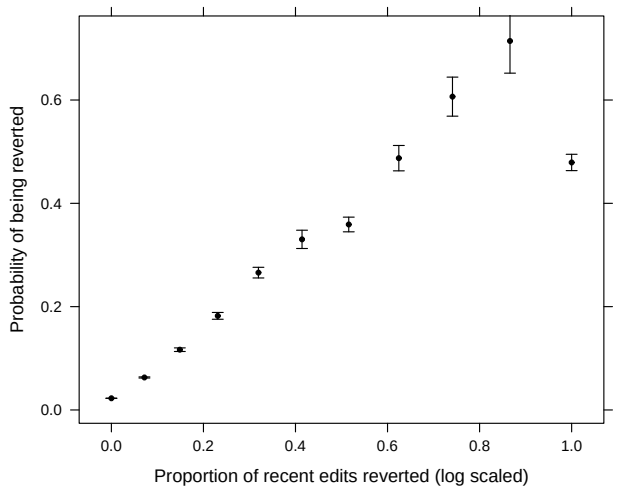

Figure 4: Probability of being reverted by the proportion of the last 20 revisions of an editor that were reverted.

flow of real-time.

In order to ensure that our results were not simply an effect of vandalism, we required a way to control for the amount of reverts that were caused directly by vandalism. To perform this normalization, we examined reverting edit comments in order to detect which edits were reverted for vandalism[8] and scaled for the amount of vandalism that we were unable to detect based on numbers discovered by a manual coding performed by Priedhorsky et al. [15]. We make the assumption that vandalism that is not detected through edit comments is distributed similarly to vandalism that is detected. Figure 3 plots the two sets of data. The top curve of points represents the probability of an edit being reverted given the average quality editors have demonstrated with their last 20 edits. The lower curve plots the points after normalizing for vandalism. Although the curve does fall with the normalization, the trend remains.

Even with the logarithmically scaled x axis, the predictive power of recent quality is centered in relatively low values. Since there are so few high values in the sample (only 21% of revisions have a recent word persistence value > 128), the metric is a powerful predictor for the majority of samples. When we ran the vandalism-controlled subsample through our model, it confirmed that recent quality continues to be a strong predictor even when the effects of vandlism are removed.

HYP *Editor Recent Reverted*. This hypothesis is interesting because it seeks to answer something very basic about our research into why work is rejected. Support for this hypothesis would answer the question, "Is the amount of reverting taking place a quality of an editor?" For example, if specific editors tend to have their work reverted because of some characteristic of themselves and not their environment, it would be reasonable to assume that editors that have a recent history of being reverted would continue to be reverted.

[8]Vandalistic reverts were detected by looking for references to vandalism in the edit comments of the reverting revision with the "d-loose" agorithm introduced by [15].

The most simplistic way to measure recent reverts would be to simply sum up the number of reverts that took place in a fixed time span. For reasons similar to those in when evaluating **HYP *Editor Recent Quality***, we decided to use the proportion of edits reverted over the last 20 edits performed by the editor as our explanatory variable. Figure 4 plots the proportions of recent revisions reverted by the probability that a subsequent revision will be reverted. As we expected, the graph shows a linear growth, indicating that the proportion of recent edits that have been reverted is a good predictor of the probability a future edit will be reverted. Our model confirms that both the proportion of recent revisions reverted for vandalism and otherwise are strong, significant predictors.

One possible cause for such high correlation is that editors who do not continue editing for long within the system are reverted frequently (as we will see in Section **HYP *Editor Experience***). If this were true, it would mean that the proportion of recent edits that have been reverted would, therefore, just be a proxy for the editor's experience. To test for this confound, we checked the correlation between an editor's experience and the proportion of their last 20 edits that are reverted. Table 2 shows that the tenure of an editor has a small, negative correlation with reverted proportions (r is -.12 and -.16 for vandalistic and non-vandalistic reverts respectively) as does the correlation with the total number of days that the editor will continue to edit (r is -.13 and -.11). This independence is confirmed by our model that shows that both the proportion of revisions recently reverted for vandalism or otherwise are powerful and significant predictors ($p < .001$) despite the effect of the total days the editor will remain active.

4.3 Indirect Editor Quality

HYP *Editor Experience*. Previous experience, as measured by previous sessions, was one of the most powerful predictors of whether an edit will be reverted or not. (See Table 1 for comparison to other explanatory variables). The power and significance of previous sessions was echoed in the amount of time since an editor began editing Wikipedia. At

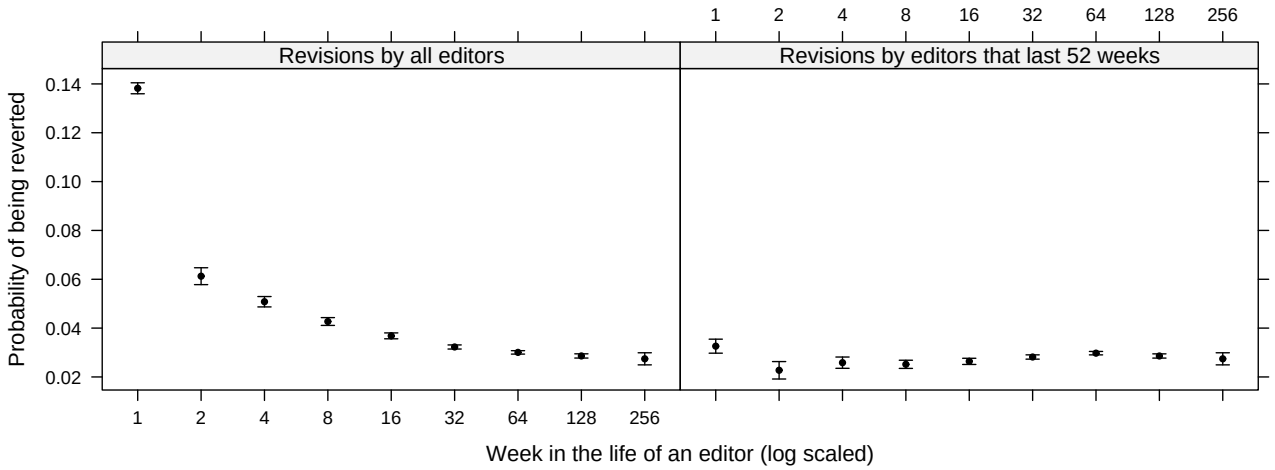

Figure 5: The probability of being reverted by weeks since editor first edited Wikipedia separated into subsets based on how long the editors will eventually survive.

first glance, this result seems to show strong support for **HYP** ***Editor Experience***.

To determine whether we were seeing the effect of learning through experience within the system, we created a subsample of editors who last at least a year in Wikipedia. Figure 5 is a side-by-side plot of the complete sample and our subsample of editors who will last. In the full sample plot, the probability of being reverted falls as we sample revisions by editors with more experience, but the subsample plot shows no appreciable fall in this probability through the life of editors. When we added the total number of days the editor would eventually be active to our model (see Table 1), tenure became an insignificant explanatory variable ($p = .37$). We saw a similar effect when controlling for the total amount of sessions an editor would eventually complete. This result suggests that the predictive power of experience is more deeply affected by a drop-out effect of highly reverted editors than any learning editors may be doing — i.e., editors don't improve as they gain experience, but instead, start out being reverted at a specific rate that predicts the amount of time they will continue editing.

Although our results support the hypothesis that an editor's level of experience is a powerful predictor of when a revision will be reverted, this analysis does not support the premise that the act of gaining experience through using the system makes editors less likely to be reverted. Our model, however, did detect a slight significant increase in the probability of being reverted with experience when we sampled only editors that would remain for at least three months. This change in the prediction supports one of the observations of Bryant et al. — that editors become more bold as they gain experience [6]. This evaluation is further supported by the slight positive correlation ($r = .03$) between the amount of time an editor has been editing and how established the words that they remove tend to be.

HYP ***Editor Policy Knowledge.*** In order to estimate knowledge of policy, we used two metrics: the number of references to policy in comments attached with edits to articles and the number of references to policy added in talk page edits. In order to differentiate between normal prose and references to policy, we used a simple regular expression that matched either "WP:<policy name>" or "Wikipedia:<policy name>".

Our analysis showed that the number of policy references that an editor has completed is not a powerful or significant predictor of when a revision will be reverted. We performed our analysis under the assumption that only those editors with knowledge of policy would reference it. Our measure only accounts for use of policy which may not be a strong proxy to knowledge of what the policy means. It could be that there is some other measure of knowledge of policy that would be a better measure, such as edits to policy pages or activity in related projects that could better identify true knowledge of policy.

4.4 Ownership

HYP ***Stepping on Toes.*** In order to gather those editors who will notice when their words are removed, we use the active editors estimate described in Section 3.3. Our hypothesis assumes that the more active editors who will notice that their words have been removed (in essence having their toes stepped on), the more likely it is that one of those editors will come back to the article to revert the change. It seemed likely to us that the number of toes stepped on could simply be a proxy for the amount of words removed by an edit. In order to ensure that this was not the case, we consulted the model and our correlation table. Since the model suggests number of active editors is independently significant ($p < .001$) and the correlation between it and the number of words removed is low ($r = .01$), the effect is independent.

Figure 6 shows the change in the probability that an edit will be reverted depending on how many active editors toes are stepped on by the edit. The figure shows linear progression of increasing probability of being reverted as the number of editors whose words were removed increases logarithmically. Note also that this is one of the of graphs that shows an expected probability as high as 0.5. In other words, depending on the number of active editors whose words are removed by the current edit, the probability of being reverted can rise 50%.

While we tested all features over many different subsets, the number of active editors with words removed was par-

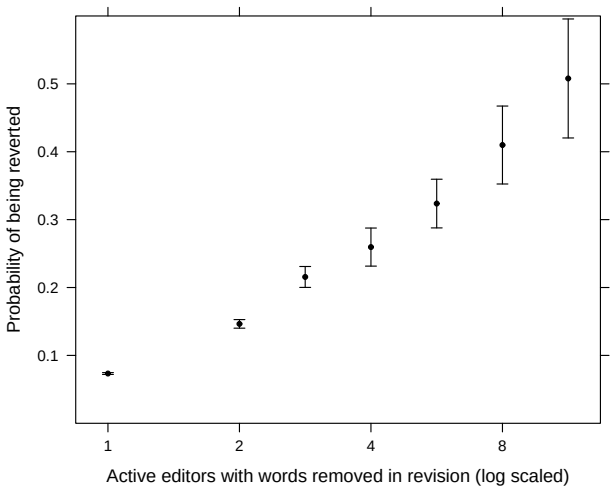

Figure 6: Probability of being reverted by the number of active editors with words removed.

ticularly interesting, because it was the only feature used in our regression model that did not lose any power for any of the subsets we tried. This suggest that the probability of a high quality, experienced, low revert proportion editors being reverted is affected in the same way by removing active editors' words as their counterparts, the low quality, inexperienced editors who are reverted often. This result strongly confirms our hypothesis.

4.5 Grouped Analysis

In order to compare the effect of our metrics against each other to ensure their independence and significance we combined them all into a logistic regression. This regression, as documented in Table 1, contains three important values for each explanatory variable and subset. The estimate (Est.) represents the change in the log odds of the response being positive (a revision being reverted) given a rise of one standard deviation of that particular explanatory variable. The standard error (SE) is the variation of the estimate. The p-value ($P(Z < |z|)$) is the probability of a variable having the estimated effect on the prediction *independently* from the other explanatory variables if there truly was no effect.

We used this model to answer questions about how our explanatory variables interact and to compare their power and utility for predicting when a particular revision will be reverted. In order to thoroughly represent our results, we've included the output from generating the model over two subsets: the complete sample and only edits by editors 90 days after they started editing. In order to come to conclusions stated in the results of the hypotheses, we performed the regression over subsets that are not listed, but simply do not have space to show them here.

Table 2 shows the correlation matrix that corresponds to the regression model described in Table 1. There are two pairs of explanatory variables that are correlated highly enough to cause concern for multicollinearity[9]. The number of days since an editor had begun editing has a high correlation with the number of session an editor had previously completed ($r = .58$) and the total days an editor will be active within Wikipedia is highly correlated with the days since an editor started editing ($r = .62$). We used the variance inflation factor of the model to determine that multicollinearity was low for all explanatory variables (< 2).

In Table 1, we can see that the significance of each of the explanatory variables persist through both subsamples with the exception of editor tenure, which only becomes significant in the old editors sample and the number of words removed by an edit that becomes insignificant. Notice that between the full sample and sample of revisions by old editors all explanatory variables fall in power with the exception of the number of active editors with words removed and the tenure of an editor. These changes suggest that as editors have been editing the Wikipedia system for a longer period of time, their history of being reverted, number of edits and removal of established words make less of a difference in their probability of being reverted. This could suggest that, although editors may not be reverted less in a significant way while they gain experience, they may ultimately be reverted for different reasons than when they were new to the system.

[9]Multicollinearity is a statistical phenomenon where two highly correlated variables cause erratic behavior in the individual estimates of a regression.

5. CONCLUSIONS

Table 3 shows the six hypotheses with a high level evaluation of our findings. The rest of this section lists the hypotheses, discusses the findings and suggests implications for future research. The order in which we present the results is different from the order in section 4 so that we may draw together sets of results that have related implications for future work.

Table 3: Tabulated conclusions by hypothesis. The right column is the level of support.

Hypothesis	Support
HYP *Removing Established Words*	Strong
HYP *Editor Recent Quality*	Strong
HYP *Editor Recent Reverted*	Strong
HYP *Editor Experience*	Mixed
HYP *Editor Policy Knowledge*	Weak
HYP *Stepping on Toes*	Strong

Our results strongly supports **HYP *Removing Established Words***. The amount of time a word has persisted in an article predicts whether an edit that removes it will be reverted. This result supports the observation by Viégas et al. of the first mover effect [19]. We also found strong support for **HYP *Stepping on Toes***, that the more active editors whose words are removed by an edit, the higher the probability will be that the edit will be reverted. The power of this feature does not depend in any way on the recent quality or experience of the editor. This result supports the supposition that editors' feelings of ownership may inappropriately lead them to discard high quality edits. One of the reasons that these results are particularly interesting is because the features on which they depend are invisible to editors.

Future research could implement an interface that makes these invisible features salient to an editor. Such an interface could have two positive effects. First, editors who are making an edit that is likely to be reverted could be coached into discussing the edit with the affected editors before continuing. Viégas et al.[20] found that a significant amount of planning occurs on article talk pages, and that this planning appears to "play a crucial role in fostering civil behavior and

Table 1: Two logistic regression coefficients and p-values. "All applicable revisions" covers all of the revisions in the sample. "Revisions by old editors" covers a revisions that were made by editors after they were 90 days old. For the discussion, statistical significance corresponds to $\alpha = 0.01$.

	All Revisions			Revision by old editors		
	Est.	SE	$P(>\|z\|)$	Est.	SE	$P(>\|z\|)$
(Intercept)	-3.512	.008	$< .0001$	-3.640	.009	$< .0001$
Total days an editor will be active (total days)	-0.098	.010	$< .0001$	-0.062	.011	$< .0001$
Recent quality (log PWRpW of last 20 edits)	-0.183	.007	$< .0001$	-0.161	.008	$< .0001$
Days since an editor began editing (current tenure)	-0.008	.009	0.3742	0.029	.010	0.0060
Previous additions to talk pages citing policy	-0.018	.012	0.1311	-0.024	.015	$< .0001$
Experience via completed sessions	-0.252	.010	$< .0001$	-0.159	.010	$< .0001$
Proportion edits recently reverted for non-vandalism	0.318	.004	$< .0001$	0.276	.004	$< .0001$
Proportion edits recently reverted for vandalism	0.217	.003	$< .0001$	0.128	.004	$< .0001$
Edits reverting other editors	0.053	.004	$< .0001$	0.049	.005	$< .0001$
Active editors w/words removed (log active editors removed)	0.231	.004	$< .0001$	0.249	.005	$< .0001$
Persistence of removed words (log PWRpW of removed words)	0.164	.006	$< .0001$	0.128	.007	$< .0001$
Number of words added by edit	0.024	.004	$< .0001$	0.023	.004	$< .0001$
Number of words removed by edit	0.006	.002	0.008	0.005	.002	0.0476
Interaction between recent quality and persistence of removed words	0.016	.005	0.002	0.019	.006	0.0013

Table 2: Correlation table of explanatory variables.

	1	2	3	4	5	6	7	8	9	10	11	12
1. Total days an editor will be active	1.0	.13	.58	.00	.40	-.13	-.12	.09	.02	-.06	-.01	.00
2. Recent quality	.13	1.0	-.07	-.03	-.22	-.11	-.09	-.17	-.13	.10	-.08	-.02
3. Days since an editor began editing	.58	-.07	1.0	.02	.62	-.12	-.16	.19	.06	.03	.00	.00
4. Previous additions to talk pages citing policy	.00	-.03	.02	1.0	.04	.00	.00	.03	.00	.01	.00	.00
5. Experience via completed sessions	.40	-.22	.62	.04	1.0	-.15	-.14	.32	.04	-.05	.02	.00
6. Edits recently reverted for non-vandalism	-.13	-.11	-.12	.00	-.15	1.0	.09	.03	.09	.10	.03	.02
7. Edits recently reverted for vandalism	-.12	-.09	-.16	.00	-.14	.09	1.0	-.01	.02	.05	.02	.01
8. Edits reverting other editors	.09	-.17	.19	.03	.32	.03	-.01	1.0	.13	.03	.02	.01
9. Active editors w/words removed	.02	-.13	.06	.00	.04	.09	.02	.13	1.0	.13	.07	.07
10.Persistence of removed words	-.06	.10	.03	.01	-.05	.10	.05	.03	.13	1.0	.00	.01
11.Number of words added by edit	-.01	-.08	.00	.00	.02	.03	.02	.02	.07	.00	1.0	.03
12.Number of words removed by edit	.00	-.02	.00	.00	.00	.02	.01	.01	.07	.01	.03	1.0

community ties". Second, previous work by Zhang and Zhu suggests that inexperienced editors are susceptible to disruptions in their intrinsic motivation when their work is edited [23]. An interface that warns inexperience editors when their work is unusually likely to be reverted might inoculate them from the demotivation.

When evaluating **HYP *Editor Recent Quality***, we found three pieces of evidence that support the assumption that word persistence (as measured by the persistent word revision per word metric) is, in fact, an approximation of the percieved quality of an editor's contributions. First, we found that articles that are edited by high word persistence editors are more likely to rise in their Wikipedia 1.0 Assessment quality rating than articles edited by lower word peristence editors. Second, the word persistence of an editor's recent work is a strong predictor of when that editor's contributions will be rejected. Third, the number of reviews a word survives is a strong predictor of whether the edit that removes the word will be reverted. This word persistence metric is convenient for future research because it can be applied to any edit in Wikipedia with information that is already publicly available. In addition to its use as a proxy for quality, future researchers might want to explore word persistence as a direct measure of the impact an editor has in Wikipedia.

HYP *Editor Recent Reverted* directly answers the question, "Is being reverted a quality of an editor?" Our results suggest that being reverted is very much a quality of an editor. However, we cannot conclude whether the reverts are because of the quality of editors' work, the characteristics of their edits (e.g. copy edits vs. content removal) or the type of articles on which they work. Future work could examine which of these characteristics best explains this phenomenon.

The amount of time editors have been active in Wikipedia and the number of sessions they have completed are powerful predictors of whether their contributions will be rejected. However, both of these variables lose their predictive power when we control for how long editors will continue to edit and how many sessions they will eventually complete. This change in predictive power occurs because editors who are frequently reverted drop out of Wikipedia quickly. Because of this dropout effect, we judge the evidence for **HYP *Editor Experience*** to be mixed since we found no evidence of a learning effect in Wikipedia editors as they gain experience despite the usefullness of experience as an explanitory variable. Future work could examine why a learning effect is not apparent among editors in Wikipedia and whether mentoring or policy changes could help to encourage editors to produce more acceptable work.

The hypothesis for which we found the least support is **HYP *Editor Policy Knowledge***. Editors who cited policy frequently were no less likely to be reverted than editors who seldom cited policy. It is important to note that

our measurements only take into account citations to policy; there may be better measures of an editor's knowledge of policy, such as surveys or tests. It is also possible that the use of policy correlates with edits to controversial content, artifically inflating the number of reverts. A measure of controversy, such as those developed by [21], can be used to test whether this result holds when controlling for controversy.

In this paper, we examined factors that seem likely to influence the probability that a contribution to Wikipedia will be rejected. Figure 1 summarizes the key factors: the quality of work removed, direct and indirect measures of the quality of the editor and feelings of ownership by other editors. Figure 1 also identifies whether each of these factors should have a positive or negative effect on the probability of rejection in an ideal peer review system. We constructed a regression model that includes key measures of all four of these factors and controls for many of the likely confounds. We observed two ways in which the Wikipedia edit process diverges from the ideal. First, neither indirect measures of editor quality had the hypothesized effect. Even experience, which has proven valuable in a wide variety of domains, does not appear to help editors avoid rejection. Second, ownership of removed content has a powerful and consistent effect on the probability of work being discarded. This result suggests that Wikipedia's review system suffers from a crucial bias: editors appear to inappropriately defend their own contributions.

Although this analysis was performed over only one informal peer review system, Wikipedia, the methods we used are generalizable to any peer review system in which common artifacts are created through collaboration and work can be discarded. Future work could use our model to look for similar trends in other peer review systems in order to determine if the relationships in which we discovered are common to all peer review systems or unique to Wikipedia.

6. ACKNOWLEDGEMENTS

This work would not have been possible without the support and assistance of our research group. This work has been supported by the National Science Foundation under grants IIS 05-34420.

7. REFERENCES

[1] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW'07*. ACM, 2007.

[2] Ahmed, Elsheikh, Stratton, Page, Adams, and Wass. Outcome of transphenoidal surgery for acromegaly and its relationship to surgical experience. *Clinical Endocrinology*, 50:561–567, May 1999.

[3] L. Argote. *Organizational Learning: Creating, Retaining, and Transferring Knowledge*. Kluwer Academic Publishers, Norwell, MA, USA, 1999.

[4] I. Beschastnikh, T. Kriplean, and D. W. McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *AAAI International Conference on Weblogs and Social Media*, 2008.

[5] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedia. *Information Visualization*, 7:34–48, 2008.

[6] S. L. Bryant, A. Forte, and A. Bruckman. Becoming Wikipedian: Transformation of participation in a collaborative online encyclopedia. In *GROUP'05*. ACM, 2005.

[7] M. T. H. Chi, R. Glaser, and E. Rees. Expertise in problem solving. Technical report, Pittsburgh Univ., PA. Learning Research and Development Center, 1981.

[8] S. Cole, J. R. Cole, and G. A. Simon. Chance and consensus in peer review. *Science*, 214(4523):881–886, 1981.

[9] A. C. Justice, M. K. Cho, M. A. Winker, J. A. Berlin, and D. Rennie. Does masking author identity improve peer review quality? *JAMA*, 280(3):240–242, July 1998.

[10] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *CSCW'08*. ACM, 2008.

[11] A. Kittur, B. Suh, B. A. Pendleton, and E. H. Chi. He says, she says: Conflict and coordination in Wikipedia. In *CHI'07*. ACM, 2007.

[12] T. Kriplean, I. Beschastnikh, and D. W. McDonald. Articulations of WikiWork: Uncovering valued work in wikipedia through barnstars. 2008.

[13] T. Kriplean, I. Beschastnikh, D. W. McDonald, and S. A. Golder. Community, consensus, coercion, control: CS*W or how policy mediates mass participation. In *GROUP'07*. ACM, 2007.

[14] A. Mockus, R. T. Fielding, and J. Herbsleb. A case study of open source software development: The Apache server. In *ICSE'00*. ACM, 2000.

[15] R. Priedhorsky, J. Chen, S. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP'07*, Sanibel Island, FLorida, USA, 2007.

[16] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. ACM Press, 2005.

[17] J. Thom-Santelli, D. R. Cosley, and G. Gay. What's mine is mine: Territoriality in collaborative authoring. In *CHI'09*. ACM, 2009.

[18] S. van Rooyen, F. Godlee, S. Evans, R. Smith, and N. Black. Effect of blinding and unmasking on the quality of peer review: A randomized trial. *JAMA*, 280(3):234–237, July 1998.

[19] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI'04*. ACM, 2004.

[20] F. B. Viégas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in Wikipedia. In *HICSS '07*, Washington, DC, USA, 2007. IEEE Computer Society.

[21] B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *WSDM'08*. ACM, 2008.

[22] J. Wales. Wikipedia sociographics. 21st Chaos Communication Congress http://ccc.de/congress/2004/fahrplan/event/59.en.html, December 2004.

[23] X. Zhang and F. Zhu. Intrinsic motivation of open content contributors: The case of Wikipedia. *Workshop on Information Systems and Economics*, 2006.