

Maintaining Efficiency in Open Production Systems at Scale  
A Case Study of Wikipedia

A Dissertation  
SUBMITTED TO THE FACULTY OF  
THE UNIVERSITY OF MINNESOTA  
BY

Aaron Halfaker

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

John Reidl & Loren Terveen

December, 2013



You see, one thing is, I can live with doubt and uncertainty and not knowing. I think it's much more interesting to live not knowing than to have answers which might be wrong.

— Richard Feynman

Dedicated to the memory of my advisor and friend, John Riedl.

1962–2013

## ABSTRACT

---

This dissertation represents an exploration of the function and failures of critical sub-systems in open production communities with Wikipedia as a case study. Specifically, I explore the nature of rejection via Wikipedia's informal, post-hoc quality control system and identify a consistent ownership bias that undermines Wikipedia's ethos of openness. I also quantify an inherent trade-off between the speed and efficiency of quality control in Wikipedia and the motivation of rejected contributors – especially new editors. I then proceed to show how Wikipedia's shifting focus on quality control and formal process has led to a dramatic decline in the rate of retention of desirable new editors that threatens the long-term viability of the project.

In light of these results, I present studies of two experimental software systems intended to explore potential solutions to this steady decrease in participation. First I draw on social learning theory to evaluate the effectiveness of a new mode of peripheral participation through reader-submitted feedback. I experimentally demonstrate effective strategies for increasing the rate of contributions without decreasing quality and argue for efficient moderation support in order to make quality control worth volunteer time spent away from editing the encyclopedia. Next, I describe the design and three month field study of a new intelligent software system intended to both efficiently support socialization practices in Wikipedia and bring visibility to the systemic problems that lead to declining newcomer retention. I show evidence that the system works in both regards: critical newcomer socialization activities are made dramatically more efficient and users of the system reflect openly on the breakdowns in Wikipedia's quality control processes.

This work has already had impact within the Wikipedia community and in directing the strategy employed by the Wikimedia Foundation in designing and evaluating new software for Wikipedia editors.

## CONTENTS

---

List of Tables	vii
List of Figures	viii
Publications	xii
<b>i BACKGROUND</b>	<b>1</b>
<b>1 INTRODUCTION</b>	<b>2</b>
1.1 Traditional vs. open production	3
1.1.1 Permission: Who is allowed to contribute?	3
1.1.2 Ownership: Who owns the content?	4
1.1.3 Governance: Who makes the rules?	5
1.1.4 So what?	5
1.2 Why study Wikipedia?	6
1.3 Functional components of Wikipedia	7
1.3.1 Production	8
1.3.2 Rule-making and governance	9
1.3.3 Enforcement	10
1.4 Wikipedia as a system	11
1.4.1 Formalization	12
1.4.2 Designing for change in socio-technical systems	12
1.4.3 Identifying and affecting emergent properties	12
1.5 Conclusion and overview	14
<b>ii CONTRIBUTIONS</b>	<b>15</b>
<b>2 HOW QUALITY, EXPERIENCE AND OWNERSHIP PREDICT REJECTION</b>	<b>16</b>
2.1 Introduction	16
2.2 Hypotheses	17
2.2.1 Quality of Content Changed	18
2.2.2 Direct Editor Quality	18
2.2.3 Indirect Editor Quality	19
2.2.4 Ownership	20
2.3 Methods	20
2.3.1 Estimating the quality of a contribution	21
2.3.2 Measuring experience	23
2.3.3 Measuring ownership and active editors	24
2.4 Results and Discussion	24
2.4.1 Quality of work	24
2.4.2 Direct editor quality	25
2.4.3 Indirect Editor Quality	28
2.4.4 Ownership	29
2.4.5 Grouped Analysis	30
2.5 Summary	32
<b>3 HOW REJECTION AFFECTS EDITORS' WORK</b>	<b>35</b>
3.1 Introduction	35
3.1.1 Research questions	35
3.1.2 Structure and contributions	37

3.2	Related Work	37
3.3	Methods	39
3.3.1	Dataset	39
3.3.2	Quantity	39
3.3.3	Quality	39
3.3.4	Boldness	40
3.3.5	Productivity	40
3.3.6	Measuring communication	41
3.3.7	Measuring changes	41
3.3.8	Notes on figures and tables	41
3.4	Results and Discussion	42
3.4.1	RQ1: How does being reverted affect the quantity of editor work?	42
3.4.2	RQ2: How does being reverted affect the quality of editor work?	46
3.4.3	RQ3: How does being reverted affect communication?	48
3.4.4	RQ4: How does experience moderate the effects of reverts on contribution?	50
3.5	Conclusions	53
3.5.1	The impact of these metrics for measuring wiki-work	54
4	THE RISE AND DECLINE OF WIKIPEDIA	55
4.1	Introduction	55
4.2	Motivation and hypotheses	57
4.2.1	Rejection of newcomers	57
4.2.2	De-personalized welcoming of newcomers	58
4.2.3	Calcification of norms against newcomers	60
4.3	Methods	61
4.3.1	First edit session	61
4.3.2	Detecting rejected contributions	62
4.3.3	Effect of rejection on retention	62
4.3.4	Newcomer quality	63
4.3.5	Tracking algorithmic tools	63
4.3.6	Conflict discussion reciprocation	64
4.3.7	Policy growth and calcification	64
4.4	Rejection & retention	65
4.4.1	Results	65
4.4.2	Discussion	67
4.5	Tool use and consequences	68
4.5.1	Results	68
4.5.2	Discussion	70
4.6	Norm formalization and calcification	71
4.6.1	Results	71
4.6.2	Discussion	72
4.7	Conclusion	73
4.7.1	Recommendations	74
5	EXPANDING PARTICIPATION AT THE PERIPHERY	75
5.1	Introduction	75
5.1.1	Participation in peer production	75
5.1.2	Supporting peripheral participation	76

5.1.3	Article feedback	77
5.2	The Article Feedback Tool	78
5.3	Research Questions	79
5.3.1	RQ1: How do different elicitations affect the volume and utility of feedback?	79
5.3.2	RQ2: How does the prominence of the elicitation affect the volume and utility of feedback?	80
5.3.3	RQ3: How does the presence of the feedback interface affect new editor conversion?	80
5.4	Methods	81
5.4.1	The article sample	81
5.4.2	Feedback utility	81
5.4.3	New editor productivity	82
5.5	Experiments and Results	83
5.5.1	RQ1: How do different elicitations affect the volume and utility of feedback?	83
5.5.2	RQ2: How does the prominence of the elicitation affect the volume and utility of feedback?	86
5.5.3	RQ3: How does the presence of the feedback interface affect new editor conversion?	89
5.6	Conclusions	92
6	SNUGGLE	95
6.1	Introduction	95
6.2	Wikipedia's socio-technical problems	96
6.3	Design strategy	97
6.4	How do we effect change?	99
6.4.1	Successor systems	99
6.5	Design of Snuggle	100
6.5.1	Participatory design process	100
6.5.2	System overview	101
6.5.3	Desirability sorting	101
6.5.4	Social literacy via traces	103
6.5.5	Social translucence	105
6.6	What is Snuggle?	106
6.6.1	Snuggle as a newcomer socialization tool	106
6.6.2	Snuggle as a critique/successor system	106
6.7	Interview study	107
6.7.1	Results & discussion	108
6.8	Conclusion	110
6.8.1	Implications for design	110
6.8.2	Implications for designers	111
iii	CONCLUSION	113
7	CONCLUSION	114
7.1	Summary	114
7.1.1	My process	114
7.2	Impact	115
7.3	Future work	117

BIBLIOGRAPHY 119

iv APPENDIX 127

Declaration 128

## LIST OF TABLES

---

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 $\alpha = 0.01$ . 31
Table 2	Correlation table of explanatory variables. 32
Table 3	Tabulated conclusions by hypothesis. The right column is the level of support. 33
Table 4	Regressions over article activity $\Delta/\sigma$ , survival and PWR/day $\Delta/\sigma$ for four weeks after the sampled edits. Characteristics of the sampled edit’s change to an article (words added, words removed, establishment of removed words) and whether it was a reverting edit itself or reverted back to by another editor are included in the regression to control for effects they could have on future work. For the discussion, statistical significance corresponds to $\alpha = .025$ . Multicollinearity was checked for using correlation between explanatory variables. All correlation coefficients are below 0.5. PWR $\Delta/\sigma$ is scaled and logged to normalize it. 43
Table 5	Dependent variable characteristics. n for all dependent variables is 684,508 after removing non-finite values. 44
Table 6	The coefficients of a logistic regression over the first edit session of two sets of randomly sampled Wikipedia users predicting survival are presented. All newcomers represents a purely random sample of registered users from Wikipedia. Desirable newcomers represents the subset of editors sampled for quality analysis that were determined to be at least acting in good-faith. 66
Table 7	The coefficients of a logistic regression over the contributions of registered editors to norm pages predicting success (i.e. not reverted) are presented. 73
Table 8	A summary of hypotheses and findings. 93

## LIST OF FIGURES

---

- Figure 1 **The English Wikipedia editor decline.** The number of active editors ( $\geq 5$  edits/month) is plotted over time for the English language Wikipedia. 7
- Figure 2 A model of how various factors should affect the probability of being reverted in an ideal system. (+) represents a positive correlation, (-) represents a negative correlation and (o) represents no correlation. 17
- Figure 3 **Persistent Word Revisions.** The word persistence values for five revisions of a sample article about apples are presented with arrows that show how words persist between revisions. Stop-words are greyed out since they are not considered in the algorithm. Revisions #3 reverts back to revision #1 and restores the word “red”. 22
- Figure 4 **Revert probability by removal of established words.** The probability of being reverted is plotted by the average persistence of words removed (as measured by word persistence). 25
- Figure 5 **Revert probability by recent word persistence.** The probability of being reverted is plotted by the average word persistence of the editor’s last 20 edits. The higher points represent the full sample while the lower were plotted after controlling for vandalism. 26
- Figure 6 **Revert probability by recent reverts.** The probability of being reverted is plotted by the proportion of the editors last 20 revisions that were reverted. 27
- Figure 7 **Revert probability by editor tenure.** The probability of being reverted is plotted by weeks since editor started editing split into subsets based on how long the editors will eventually survive. 28
- Figure 8 **Revert probability by toes stepped on.** The probability of being reverted is plotted by the number of active editors with words removed. 30
- Figure 9 Hidden variables of editor activity are connected to the metrics that were used as proxies in the analysis and are divided by the research questions they are used to explore. Metrics obtained from 2008 data are signified with an “\*”. An arrow from A to B means “A is used as a proxy for B”. The dotted line between boldness and quality represents the confounding effect described in section 3.3.4. RQ4 does not rely on hidden variables. 38
- Figure 10 **Article activity  $\Delta/\sigma$ .** For four weeks after a sampled edit, the change in article activity is reported. Reverted edits are split by whether the reverting editor was registered or anonymous. A control group of similar editors who were **not** reverted is included for comparison. 42

- Figure 11 **Reverts per revision  $\Delta$ .** For the four weeks immediately after a sampled edit, the change in quality of work as reverts/revision is reported. A control group of similar editors who were **not** reverted is included for comparison. 47
- Figure 12 **Changes in boldness.** For the four weeks immediately after a sampled edit, the change in the boldness of work is reported via two metrics: words changed/revision and establishment of removed words. A control group of similar editors who were **not** reverted is included for comparison. 48
- Figure 13 **PWR/day  $\Delta/\sigma$ .** For the four weeks immediately after a sampled edit, the change in productivity is reported as the controlled PWR/day delta. A control group of similar editors who were **not** reverted is included for comparison. 49
- Figure 14 **Communication activity  $\Delta/\sigma$ .** For the four weeks immediately after a sampled edit made by surviving editors (defined in Section 3.4.1), the change in Article\_talk and User\_talk communication activity is reported. A control group of editors with similarly distributed tenure who were **not** reverted is included for comparison. 50
- Figure 15 **Article activity  $\Delta/\sigma$  by reverted editor tenure.** For the four weeks immediately after a sampled edit, the change in article activity is reported for newbies and old-timers. A control group of similar editors who were **not** reverted is included in each graph for comparison. 51
- Figure 16 **Article activity  $\Delta/\sigma$  by reverting editor tenure.** For the four weeks immediately after a sampled edit, the change in article activity is reported. Reverted edits are split by whether the reverting editor was a newbie or old-timer. A control group of similar editors who were **not** reverted is included for comparison. 52
- Figure 17 **The English Wikipedia editor decline.** The number of active editors ( $\geq 5$  edits/month) is plotted over time for the English language Wikipedia. 55
- Figure 18 **Quality of newcomers over time.** The proportion of editors falling into the two good-faith quality categories is plotted over time. 66
- Figure 19 **Reverts of desirable newcomer contributions over time.** The proportion of good (“good-faith” & “golden” combined) newcomers with at least one reverted first session edit is plotted over time. 67
- Figure 20 **Survival of desirable newcomers over time.** The proportion of surviving good (“good-faith” & “golden” combined) newcomers is plotted over time. 68
- Figure 21 **Use of algorithmic tools to reject newcomers edits.** The proportion of rejected first session contributions is plotted for newcomers by the mechanism used for rejection over time. 69
- Figure 22 **Use of algorithmic tools to reject newcomers edits.** The proportion of rejected first session contributions is plotted for newcomers by the mechanism used for rejection over time. 70

- Figure 23 **Rate of automated reverts for desirable newcomers.** The proportion of reverted desirable newcomers (“good-faith” & “golden” combined) who were reverted using algorithmic tools is plotted over time. 71
- Figure 24 **Norm page growth over time.** The change to overall size of the three norm types is plotted by year. 72
- Figure 25 **The Article Feedback Tool’s interface components.** The components of the AFT interface are called out from Wikipedia’s article viewing interface. An article on Kim Manners, one of the randomly sampled articles, is loaded. #1-3 represent different versions of the article feedback forms. #4 represents the edit invitation form, a request for the reader to try editing the page. A and E represent links inserted into the page to direct readers to the feedback form. 78
- Figure 26 **The Feedback Evaluation System.** FES, the interface used by ikipedians to evaluate the usefulness and type of feedback submitted is resented with annotations for the three main components. 82
- Figure 27 **Quantity of feedback by experiment.** The median daily feedback submissions per day is plotted for the last two weeks of the experiment with box limits at the 25% - 75% quantiles and bar limits at the most extreme observed values. 84
- Figure 28 (left) The proportion of feedback submitted is plotted by intention as determined by at least one Wikipedian. (right) The proportion of useful feedback is plotted for each type. These plots draw aggregate proportions from feedback submitted through all three experimental conditions. 85
- Figure 29 **Volume of feedback by experiment.** The raw amount of feedback submitted via the three experimental conditions is plotted by whether the feedback was submitted via the prominent link or via the form at end of articles. 87
- Figure 30 **Utility of feedback by origin.** The proportion of useful feedback is plotted for each condition by the origin from which it was submitted with standard error bars. Note that the large error bars around the proportion of useful feedback submitted via the link in the 1A condition is due to the small amount of feedback sampled in that condition ( $n = 18$ ). 88
- Figure 31 **New users by origin.** The number of new editors is plotted and stacked by the origin of their first edit. “edit link” refers to the standard vector for accessing the edit pane. “invitation” refers to form 4 which invites the user to make an edit. 89
- Figure 32 **Proportion of productive new users by origin.** The proportion of new editors who made at least one productive contribution in their first week is plotted by the origin of their first edit for the three experimental conditions. 90
- Figure 33 **The cost of non-productivity.** The proportion of newcomer revisions reverted for each experimental condition by how the revision was reverted. 91

- Figure 34 **Snuggle’s user browser.** A screenshot of the Snuggle user browser is presented with UI elements called out. The user dossier for “Noorjahanbithi” is selected. An edit in the *interactive graph* is selected and information about the edit is presented. (a) The (unexpanded) categorization menu, (b) The (unexpanded) wiki action menu (see section 6.5.4.2), (c) Tabs for accessing lists of categorized users, and (d) Talk page icons representing socially relevant traces (see section 6.5.4.1). 101
- Figure 35 (left) Histograms of the frequency of STiki scores are plotted for the training set newcomers’ edits with expectation maximization fits of beta distributions overlaid. (right) The receiver operating characteristic of the desirability ratio of newcomers from the test set is plotted. 102
- Figure 36 **Wiki actions menu.** A screenshot of the “wiki action menu” is presented with a the message sending functionality selected and a test message written. Note the preview on the right side specifies which page the message will be appended to. 104
- Figure 37 **The recent activity feed.** A screenshot of Snuggle’s recent activity list is presented. 105

## PUBLICATIONS

---

Many ideas and figures presented in this dissertation have appeared previously in the publications listed below. While I performed the work for these publications in collaboration with others, only the writing, analysis and results of my own efforts have been included in this dissertation.

- Halfaker, A., Kittur, A., Kraut, R. E., & Riedl, J. A Jury of Your Peers: Quality, Experience and Ownership in Wikipedia. *WikiSym'09* ACM, New York, NY, USA, (pp. 15:1–10).
- Halfaker, A., Song, B. D., Stuart, D. A., Kittur, A., & Riedl, J. (2011). NICE: Social translucence through UI intervention, *WikiSym'11* ACM, New York, NY, USA, (pp. 101–104).
- Halfaker, A., Kittur, A., & Riedl J. (2011). Don't Bite the Newbies: How reverts affect the quantity and quality of Wikipedia work, *WikiSym'11* ACM, New York, NY, USA, (pp. 163–172).
- Halfaker, A., Keyes, O., & Taraborelli, D. (2013). Making Peripheral Participation Legitimate: Reader engagement experiments in Wikipedia, *CSCW'13* ACM, New York, NY, USA, (pp. 849–860).
- Halfaker, A., Geiger, R.S., & Terveen, L. (in press). Snuggle: Designing for efficient socialization and ideological critique, *CHI'14* ACM, New York, NY, USA.

Part I

BACKGROUND

# 1 INTRODUCTION

---

The open collaboration process as supported by information technology has emerged as a powerful and effective form of content production. The so-called "Web 2.0" generation of user-generated content systems distribute production work by blurring the line between consumer and producer. While many of these systems reflect the traditional publishing practices of individual production, ownership, and control (e.g. YouTube<sup>1</sup> and LiveJournal<sup>2</sup>), open production communities (e.g. Wikipedia<sup>3</sup> and OpenStreetMap<sup>4</sup>) that overturn these traditional practices (by distributing ownership and control as well) have shown the potential to produce immense, valuable resources. Despite this apparent success, the characteristics of these open production communities are poorly understood and have thus been the subject of focused study over the last decade.

In my work as a doctoral candidate, I've contributed to the body of work exploring the structure and function of these open production systems. I've focused my work on what I saw as one of the most essential and also most poorly understood aspects of open production: quality control. Since open systems like Wikipedia allow changes to the live version of the encyclopedia to be made at any time by any internet user<sup>5</sup>, naive intuition suggests that such a system could be trivially overrun by deviant users or would otherwise be full of nonsense. The fact that Wikipedia has attained a high level of quality and has risen to be one of the most frequently visited websites based on the efforts of volunteer internet users despite this openness to attack should raise questions from any investigator.

In this dissertation, I'll describe a series of studies that explore and attempt to influence the behavioral patterns of Wikipedia's readers and volunteer editors as they participate in the system's emergent, post-hoc, informal peer-review based quality control system. My analysis begins with answering very basic questions about how quality control works in Wikipedia. For example, in chapter 2, I'll answer the question, "What types of contributions are rejected in Wikipedia?" While exploring this phenomenon, I discovered that the practice of quality control in Wikipedia is tightly coupled with the motivation of new contributors to continue volunteering their time and energy (see chapter 3). Given that the future of Wikipedia depends on the motivation of volunteers to contribute to the encyclopedia, I've also explored the relationship

---

<sup>1</sup> <http://youtube.com>, a popular video sharing website

<sup>2</sup> <http://livejournal.com>, a popular blogging service

<sup>3</sup> <http://wikipedia.org>, a popular open, online encyclopedia

<sup>4</sup> <http://openstreetmap.org>, a popular geographic wiki

<sup>5</sup> Some practical limitations apply. A tiny proportion of articles in Wikipedia are "protected" from edits by unvetted newcomers due to their controversial nature, and of course, users who regularly cause damage will be blocked from editing entirely.

between motivation and quality control as it affects the efficiency of Wikipedia as a whole (formalized in section 1.4; see chapter 4 for relevant analysis).

In the next sections, I'll contrast traditional and open production models by discussing a set of novel problems that open production systems like Wikipedia introduce around trust, ownership, governance and enforcement. Next, I'll motivate the use of the English Wikipedia as a case study for exploration of these problems and describe Wikipedia's emergent subsystems that effectively solve these problems without centralized control. Then, I'll explain how I formalize the activities of Wikipedians (Wikipedia editors) in the context of Wikipedia as a *system* for turning the attention of volunteers into a high quality encyclopedia. Finally, I'll conclude with an overview of the results presented in the substantive chapters of this dissertation.

## 1.1 TRADITIONAL VS. OPEN PRODUCTION

Through the use of a suite of technologies built for the internet, open production communities have inverted the traditional publishing model in order allow information consumers to take part in (or even ownership of) production itself. While removing barriers to participation dramatically boosts the number of man-hours that can be dedicated to production, a novel set of problems are introduced. In this section, differences between traditional publishing, open source software and Wikipedia's open publishing model will be described with examples along three dimensions: permission, ownership and governance.

### 1.1.1 *Permission: Who is allowed to contribute?*

Under the traditional publishing model, a document has an owner and that owner is allowed to directly contribute to the document. While others may be invited to comment or make suggestions, production ultimately falls on the owner. Once an artifact is published, it is rarely updated, and if so, only through highly structured mechanisms by individuals who are given permission or explicitly invited before-the-fact. Many online user generated content communities have adopted this model. For example, YouTube and LiveJournal allow individuals to publish videos and blog entries respectively. In these systems, non-authoring users are merely allowed to comment on and vote for their preferred content.

Open source software systems take a step away from this highly constrained, individualistic model. In most open source software systems, a trusted group of "committers" are allowed to make changes to a shared collection of source code. While most trusted groups consist of only one committer [76], the most successful projects tend to gather small groups of trusted contributors who are allowed to directly change the product (source code) without permission. Untrusted contributors are allowed to submit "patches", proposed changes to the source code that can be algorithmically ap-

plied. These patches must be vetted by a trusted committer before they actually affect the software product.

Many open source software projects have developed structured processes by which developers may be granted commit rights. This process often involves a demonstration of both competency and good-faith through peripheral participation in the project (e.g. filing bugs in a tracker and participating in discussions on a mailing list) [18] For example, MediaWiki, the open source wiki software on which Wikipedia runs, is maintained by a group of tens of source code committers that were formally vetted through an application process<sup>6</sup>. By extending production to many contributors, but filtering them through a vetting process, open source software communities are able to increase the available man-hours while leveraging trust to preserve quality.

Open production systems like Wikipedia, a system which is often held up as a synecdoche for open production, drop the constraint on trust entirely by opening up the product to contributions from nearly anyone. In these systems, trust is granted without any formal vetting, but it may be revoked for contributors who show that they should not be trusted. By relaxing constraints imposed by trust-based systems, Wikipedia accepts many more contributions by readers-turned-contributors than more restrictive systems but also opens itself up to vandalism and other forms of damage.

While it may be counterintuitive, this open extension of trust seems to be well founded. My research has quantified the surprising fact that the majority of potential contributors will operate in good-faith even when contributing anonymously (see chapters 4 and 5). While critics claim that extending trust in this way comes at great cost due to the damage that is caused by less than trustworthy individuals [57], it's become apparent that Wikipedia's structure of implicit trust has been highly successful in producing a high quality [33] and important [54] resource.

### 1.1.2 *Ownership: Who owns the content?*

Under the traditional publishing model, ownership is clear and restrictive: the author of an artifact is the owner of the artifact. Online systems based on the traditional publishing model generally make use of software restrictions to enforce ownership of a document and copyright/licensing regulations to prevent unauthorized copy.

Open source software and open production systems spread ownership over their respective artifacts generally by providing the means to make copies and redistribute them. In this way open source software and open production systems like Wikipedia are quite similar. Until 2010, Wikipedia's content was licensed under the GNU Public License, a copyright license designed to ensure that derivative works of open source software would also be openly owned. Since 2010, Wikipedia has switched to the Creative Commons CC-BY-SA<sup>7</sup> license which asserts a similar set of requirements for derivative works but was deemed more appropriate for the written medium.

<sup>6</sup> [http://www.mediawiki.org/wiki/How\\_to\\_become\\_a\\_MediaWiki\\_hacker](http://www.mediawiki.org/wiki/How_to_become_a_MediaWiki_hacker)

<sup>7</sup> <http://creativecommons.org/licenses/by-sa/3.0/>

This open ownership model affords an interesting opportunity and challenge for these systems. While the combination of open participation and open ownership means that there is a larger audience of potential contributors due to the essentially free nature of the product, it also makes implementing a production incentive based on reimbursement difficult. Social Loafing theory would suggest that, without reimbursement in exchange for contributions, the requisite motivation to support production should be missing [40]. However, Wikipedia and open source software have proven an exception to this rule. Theories of motivation in social contexts like the Collective Effort Model seek to describe their success by introducing group goals and group identity to the decision making structure [41]. Under this framework, the group ownership model lends itself naturally to motivation structure of group goals and identity.

### 1.1.3 *Governance: Who makes the rules?*

Governing the production, use and distribution of traditionally published products is logistically straightforward. User-generated content communities that support the traditional publication model of individual ownership generally have an individual or offline organization responsible for acting in the role of a benevolent dictator (centralized authority) who makes all of the important governance decisions for a community. These decisions are often enforced by the allowances and restrictions of the software.

However, in an open production community like Wikipedia, there is no central authority, so governance strategies that do not require such an authority are needed. But which governance strategies work in this egalitarian context? Research into mechanisms of commons based governance has gained substantial ground in this space [71] and has been applied, with much success, to explain Wikipedia's means of governance and decision making [25, 30, 6]. Wikipedia is an interesting case of governance and enforcement issues since most editors of the encyclopedia do not have direct access to the software; most enforcement must be accomplished by agreements between individuals and within the affordances of the software before the rule was articulated.

An interesting case study in collective decision-making in Wikipedia is the policy for *consensus*<sup>8</sup> on which all important group decisions are based. In Wikipedia, consensus doesn't necessarily mean that everyone who participates in a discussion agrees, but instead, it represents a subjective assessment of "all editors legitimate concerns". While this may seem difficult to operationalize, it seems to work to settle most disputes in practice.

### 1.1.4 *So what?*

While managing content within online communities that support traditional models of publication is well understood, we have yet to build a clear understanding of how to solve problems of production, ownership and governance in open production systems.

<sup>8</sup> <http://en.wikipedia.org/wiki/Wikipedia:Consensus>

In other words, open production systems represent a novel space to explore problems and solutions similar to those of common goods maintenance in offline contexts.

## 1.2 WHY STUDY WIKIPEDIA?

My work focuses on the English Wikipedia as a case study for open production systems. Wikipedia is an excellent example for analysis because (1) as described in the previous section, its publishing model represents an extreme departure from traditional publishing in an online context (2) the system is a highly successful example of online social production with many novel open problems to explore, and (3) a complete log of 13 years of Wikipedian editing activities, discussions and policy negotiations is publicly available for analysis.

Since the online encyclopedia's inception in 2001, the encyclopedia and the community of volunteer editors who construct it has grown massively. As of Aug. 2013, the encyclopedia contains over 4 million articles<sup>9</sup>. Wikipedia has risen to the top of web search engine results[79] and has all but rendered traditional encyclopedias obsolete due to its wide, cheap distribution via the internet and its surprisingly high quality content [33]. Despite this success, the mechanisms by which Wikipedia actually *works* are still poorly understood and the subject of much study. As will become apparent in the next section, the scholarly research exploring Wikipedia has provided a breadth of analysis of Wikipedia's subsystems that has afforded me the ability to experiment within a well-documented environment.

In Wikipedia, nearly all interactions between editors and the articles they edit are captured via log data that is made publicly available by the Wikimedia Foundation<sup>10</sup>. The availability of this data has proven immensely useful for a wide range of analysis – from spam & vandalism detection algorithms [98] to strategies of collaborative governance [48, 24, 6] and automated task assignment via collaborative filtering [15]. Further, through my efforts working with the Wikimedia Foundation and building support for research within the community, the MediaWiki software on which Wikipedia runs has opened (in limited ways) to controlled experimentation.

Most important to the focus of my studies, Wikipedia is an example of a complex system with emergent subsystems and no central authority. From the interactions of practically anonymous internet users on the wiki-platform, an amazing human resource has emerged despite seemingly well-reasoned skepticism. Yet the process and nature of the of this emergence is poorly understood. A common quote (sadly difficult to attribute) that's repeated among Wikipedians sums up the problem nicely.

*Wikipedia works in practice, but not in theory.*

The goal leading my work as a doctoral student is to make substantial progress towards building understanding of how Wikipedia works and how such systems may be grown and adapted for new contexts.

<sup>9</sup> <http://stats.wikimedia.org/wikimedia/animations/growth/index.html>

<sup>10</sup> [http://meta.wikimedia.org/wiki/Data\\_dumps](http://meta.wikimedia.org/wiki/Data_dumps)

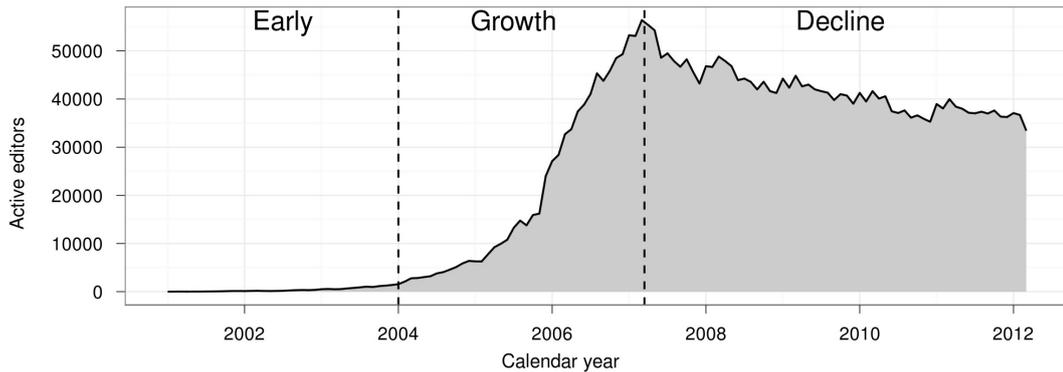


Figure 1: **The English Wikipedia editor decline.** The number of active editors ( $\geq 5$  edits/-month) is plotted over time for the English language Wikipedia.

There are also many ways in which Wikipedia does not appear to be working, in practice. The Wikipedia community is currently suffering from a potentially serious and poorly understood problem: an abrupt decline in the pool of volunteer contributors. As figure 1 suggests, the number of active volunteer editors in Wikipedia abruptly transitioned from exponential growth to a steady decline in 2007. Recent research has implicated declining newcomer retention as the source of this overall decline in active contributors [84, 99]. While this issue has been raised to the top of the Wikimedia Foundation’s strategic priorities<sup>11</sup>, my recent analysis suggests that no substantial improvement has been made since the 2007 shift in newcomer retention rates.

Since neither the research literature, nor the community members have found a viable solution to this problem – or even an empirically validated explanation of its cause – I saw exploring this problem as an opportunity to build theory about open collaborative participation and human computer interaction in socio-technical systems. Chapter 4 describes a collection of analyses and models that strongly suggest that the reason for the decline is related to Wikipedia’s quality control mechanism, and in chapter 6, I describe a software system intended to provide Wikipedians with the means to reverse the systemic problems that lead to this negative trend.

### 1.3 FUNCTIONAL COMPONENTS OF WIKIPEDIA

In section 1.1, I described a set of ways in which open production systems like Wikipedia differ from traditional publication and described a set of problems that must be solved in order for such systems take advantage of the benefit of openness: a massively increased labor force. For each of these problems, a set of functional sub-components can be identified within Wikipedia that have emerged from the work practices of Wikipedians and their use of the wiki software. In this section, I’ll summa-

<sup>11</sup> [http://wikimediafoundation.org/wiki/Wikimedia\\_Movement\\_Strategic\\_Plan\\_Summary](http://wikimediafoundation.org/wiki/Wikimedia_Movement_Strategic_Plan_Summary)

size related work describing these functional components and describe how they have scaled with the growth of Wikipedia into a massive online community.

### 1.3.1 *Production*

The system of task allocation and workload distribution in Wikipedia is highly decentralized and naturally scalable since it relies on individuals to be self-directed.

**Task allocation.** By browsing and reading the encyclopedia, readers (potential contributors given the permission model) direct themselves to the encyclopedia content that matches their interests. Assuming that readers will tend to be most likely to read about subjects that they have some knowledge about, they're also browsing the articles they could most easily contribute to.

The natural process by which readers identify opportunities to contribute that align with their interests by simply reading the encyclopedia minimizes the amount of effort necessary to contribute. This pattern bears a striking similarity to Eric Raymond's formulation of "Linus' Law" that states, "given enough eyeballs, all bugs are shallow" [76]. In the context of Wikipedia, if enough readers see an article, all potential contributions to an article will take minimum effort for someone. Notably, previous work has explored recommendation as a means to increase the rate at which editors identify such high efficiency tasks. SuggestBot<sup>12</sup> helps editors find articles in their area of interest that need work [15].

The Collect Effort Model gives us a framework for understanding when an individual's motivations can overcome the tendency towards socially loafing [41]. If a reader perceives that the value of making a contribution (to themselves and the group) outweighs the cost (which is low per the previous paragraph), Wikipedia erects no technical barrier to prevent them from doing so.

**Workload distribution.** Similar to other online communities that support user-generated content [100], Wikipedia is primarily constructed by a small group of highly prolific editors who adopt many roles in the system [96]. A much larger group of more casual editors who contribute very little per-person represent the rest of the editing community [74]. Wikipedia has historically benefitted from the complimentary work styles of these different sides of the contribution investment spectrum. There are many types of high investment work (e.g. refactoring category structure, writing policy, and authoring new content) and low investment work (e.g. copy editing, minor additions, and rewording). At least as far as investment is concerned, this allows new editors to follow a well documented pattern of peripheral participation [55] while they evaluate the community and decide whether to make a larger investment [9, 73].

---

<sup>12</sup> <http://en.wikipedia.org/wiki/User:SuggestBot>

### 1.3.2 Rule-making and governance

Recent research has documented Wikipedia's decentralized pattern of articulating and enforcing rules. As theory developed in offline contexts predicts<sup>[71]</sup>, this process is highly decentralized and constructed to meet the needs of local concerns <sup>[25]</sup>. This property allows a form of crowd governance that both solves the necessary problems of this functional component of the system and scales naturally to the size of the community.

Social norms that dictate behavior in Wikipedia emerge through interactions at many levels. An interesting example of the use of norms in Wikipedia is expressed in the structure of *talk pages*. Unlike other online discussion systems, the MediaWiki software used by Wikipedia does not explicitly support a topic and reply structure for discussions. Instead, discussions in Wikipedia take place through the collaborative editing of a shared document. The norms of behavior on these talk pages dictate a regular structure and format. Technically, any discussant (or even a random passerby) could change the wording of any other discussant, but the public nature of the change logs and strong norms keep discussions structured and ensure that the editors will leave each others' comments alone<sup>13</sup>

Wikipedians have adopted the practice of codifying emergent norms (e.g. *talk page structure*<sup>14</sup>) into "policies" and "guidelines", the equivalent of Wikipedia laws that can be used to educate, implicate or vindicate potential offenders and justify sanctions taken against them <sup>[6]</sup>. The process of converting social norms into explicit rules that can be cited to address local concerns has been shown to scale naturally with the growth of the system <sup>[25]</sup>. However, some of my more recent work (covered in chapter 4) suggests that while Wikipedia has matured this norm articulation processes has calcified; this has negative consequences for the modification and creation rules needed to solve new problems.

To settle disagreements, Wikipedians have adopted and refined the notion of "consensus". As the policy on consensus states<sup>15</sup>:

Consensus on Wikipedia does not mean unanimity (which, although an ideal result, is not always achievable); nor is it the result of a vote. Decision-making involves an effort to incorporate all editors' legitimate concerns, while respecting Wikipedia's norms.

While this decision making strategy is subjective and therefore difficult to apply consistently, it apparently works in practice as the concept has been successfully used to make decisions in Wikipedia for over a decade. More importantly, since "consensus" only requires the editors presently involved in a conversation to agree, no central authority is necessary, and therefore decisions can be made by any group of discussants.

<sup>13</sup> Although the structure of talk page discussions is a fascinating emergent pattern of Wikipedia norms circumventing the need for technical support, this property of talk pages makes analysis of discussion patterns very difficult <sup>[51]</sup>.

<sup>14</sup> <http://en.wikipedia.org/wiki/WP:TALK>

<sup>15</sup> <http://en.wikipedia.org/wiki/Wikipedia:Consensus>

This property allows the decision making process to scale naturally to the number of potential discussions [24]. For controversial discussions that fail to reach consensus, Wikipedians have constructed a tiered moderation structure that starts with a *request for comment* and can be escalated to a body of elected editors called the “Arbitration Committee” which has the ultimate authority to make decisions in disputes.

### 1.3.3 Enforcement

Rule enforcement in Wikipedia is also a decentralized process that has emerged from norms developed by the community.

The application of rules via citations to policies and guidelines in discussions has been well documented by Kriplean et al.[48] Once norms are formalized into policies and guidelines (as described above) any Wikipedia editor may make use of these rules in discussions by providing a link to them. Beschastnikh et al.’s work shows that new policies and guidelines are first adopted by administrators, the inner circle of Wikipedia [6]. After a brief period of time, the adoption spreads to highly active editors and eventually to casual editors and newcomers. This process of rule use diffusion throughout the Wikipedian community constitutes a distributed enforcement and re-articulation pattern whereby any editor may choose to apply a rule during a discussion and such rules may be re-vetted in the local circumstances of their application.

One particular space of rule enforcement dominates concerns about the quality and viability of an openly edited online encyclopedia: counter-vandalism. If the encyclopedia is open to changes by anybody, what stops deviant users from damaging the content and ruining both the value of the content and the encyclopedia’s credibility?

Research of Wikipedia’s informal review system suggests that Wikipedia maintains high quality despite its vulnerability because of the distributed, informal peer review performed by experienced Wikipedia editors [82]. Several strategies of distributed damage detection and control have emerged from the Wikipedia editor community. Since the MediaWiki software keeps a complete history of the revisions of all pages, damage can be removed by simply restoring the last good revision of a page. This operation of restoration is commonly referred to as a *revert*.

There are also several ways that editors can monitor changes to articles to identify damaging contributions, but unlike a more traditional publishing model, rejection of such damage must always happen after-the-fact. That is, damaging edits go live immediately, meaning that the wiki is left in a damaged state for some period of time until the damage can be reverted. A common way to operationalize the cost of vandalism in Wikipedia is by examining this *time to revert* [74, 45, 28].

When Wikipedia was young and there were few contributors, editors could make use of two MediaWiki features to efficiently monitor damage: the recent changes list<sup>16</sup> and watchlists<sup>17</sup>. However, as Wikipedia gained popularity and the pool of editors

<sup>16</sup> A MediaWiki feature that displays a list of the most recent edits that have been performed across the wiki

<sup>17</sup> A MediaWiki feature that allows editors to monitor changes to a set of article they are interested in

began to grow exponentially, the number of new edits to review grew beyond the scale that editors could practically manage with recentchanges and watchlists alone<sup>18</sup>.

Out of the pressure to quickly review this massive amount of edits for damage and ban deviant contributors, editors developed strategies that allowed them to multiply their damage control power. Wikipedians developed partially and fully-automated software tools and a distributed user tagging system[30]. Algorithmic tools in the form of fully-automated robots<sup>19</sup> and semi-automated human-computation tools<sup>20</sup> use machine learning and other inferential techniques to revert vandalism automatically or bring potential vandalism to the attention of editors. A user tagging system lets vandal-fighting editors draw administrator attention to contributors who have a recent history of damaging articles by tagging those contributors with a “warning template” [30]. This system of tags allows human editors and robots to independently contribute to a *rap sheet* of an editor’s activities that will attract the attention of administrators<sup>21</sup> Warning templates draw the attention of administrators who can use their authority to ban users who’ve made it clear that they will continue behave badly.

Through the combination of efficient vandalism detection and distributed coordination via the tagging structure, Wikipedians are able to keep the encyclopedia nearly vandalism free[74, 45]. Some of my other work in collaboration with R. S. Geiger (not featured in this dissertation) has shown that, when part of this system fails (e.g ClueBot NG went down for a few substantial time periods in 2011), the median *time to revert* rises substantially, but also that the system will re-adjust to the loss and all of the damage will eventually be removed [28].

#### 1.4 WIKIPEDIA AS A SYSTEM

In my work, I sought to ask research questions about Wikipedia at the “system-level” – concerns that are general to the survival and efficiency of Wikipedia as a whole. Looking at Wikipedia as a system structures my work in a few useful ways. Namely, Wikipedia has a goal: to construct a complete, high quality information resource and distribute this resource as widely as possible. In the context of this goal, we can formalize Wikipedia’s inputs (volunteer editors’ attention) and outputs (high quality encyclopedic content). Further, we can discuss efficiencies of Wikipedia for converting the available input resources into output as a whole system as well as discuss its subsystems (discussed as “functional components” in section 1.3).

<sup>18</sup> During peak times (evening GMT, morning EST), 2-3 article edits are saved every second.

<sup>19</sup> For example, ClueBot NG: [http://en.wikipedia.org/wiki/User:ClueBot\\_NG](http://en.wikipedia.org/wiki/User:ClueBot_NG)

<sup>20</sup> For example, Huggle: <http://en.wikipedia.org/wiki/Wikipedia:Huggle>

<sup>21</sup> Administrators are Wikipedia editors who have applied for additional permissions, such as the ability to block users from contributing, through a consensus driven nomination process.

### 1.4.1 Formalization

Within the Wikipedia's encyclopedia generation and distribution system, the attention and effort of contributing individuals are consumed in the process of producing a high quality encyclopedia. When looking at Wikipedia this way, we can formalize Wikipedia's productivity (equation 1) and efficiency (equation 2).

$$\text{productivity} = \text{content quantity} * \text{content quality} \quad (1)$$

$$\text{efficiency} = \frac{\text{productivity}}{\text{available human attention}} \quad (2)$$

It is my assertion that, when Wikipedia is a *healthy* system, we'll see evidence of high system efficiency.

As described in the previous section, several robust subsystems have emerged within Wikipedia that manage production, rule formulation and enforcement. In order to understand how Wikipedia's subsystems work, we need to start from an understanding of how such subsystems emerge without central authority

### 1.4.2 Designing for change in socio-technical systems

Complex adaptive systems theory provides a framework for understanding how robust, near-optimal systems emerge with no central direction. Miller describes the adaptation in social communities as a process by which the decisions made by individual members of a community within their local context will gather, propagate, maintain, or silence patterns within the larger group [60]. This process is seen as adaptive when the resulting system-level effect resembles Adam Smith's *invisible hand* whereby individual community members behave as a collective to serve the community's needs without a central authority providing direction[80].

In my work, I apply the complex systems way of thinking about emergent properties to explore how the decisions that Wikipedia's editors make on an individual, local basis result in emergent system-level properties that have implications for the overall efficiency of the system. In chapter 6, I'll draw on this way of thinking about emergent properties to design software to extend and correct some of the most concerning emergent properties identified in chapter 4.

### 1.4.3 Identifying and affecting emergent properties

Neither identifying nor affecting the source of emergent patterns complex systems has been well documented. However, there are two straightforward premises that serve as base assumptions for my reasoning:

1. Properties of complex systems emerge from the individual decisions and interactions between individuals in the system, and therefore, a system level effect can be directly understood and modified via individuals' behavior and their interactions.
2. Two potential routes of changing individual behavior exist: direct change to individuals thought or change to the environment in which individuals interact.

While reprogramming a person is still (thankfully) the realm of Sci-Fi for the time being, an online community whose environment is constructed via computer software provides a unique opportunity to change the environment in which individuals interact. For example, work by Erickson & Kellogg on *social translucence* represents a set of practical examples of how surfacing socially relevant cues in user's local environment has the power to direct behavior [20].

Affecting an individual's environment in non-virtual spaces is substantially more difficult than in digital spaces. Physical mass communication is expensive due to printing or network costs and imperfect due to competition for attention and visual space. Further, physical environments often force individuals into certain behavioral patterns. For example, occupying some spaces or performing some actions can be difficult, painful or impossible for human bodies. This makes experimentation with real world crowds difficult and expensive process.

Software system designers have immense power in the virtual environments they create since the space in which agents operate is made of software. Since software is trivial change, having a direct effect on an individual's environment becomes feasible. The environment itself can be experimented with directly and cheaply. New spaces for interaction, new information tools and software based restrictions can be imposed quickly, effectively and on a random sample of users.

However, with great power comes great responsibility [56]. Changes to shared virtual environments can have massive and sometimes destructive effects on the viability of online social systems [26], so an understanding of the norms and work patterns of the community under study is essential so as to not disrupt the complex patterns that form healthy functional components of the system [? ].

In contemporary HCI, there are two general approaches to designing software to modify or amplify behavior(c.f. [36]). Second-wave HCI seeks to design efficient interfaces and systems to support desirable social tasks, such as a tool to support Wikipedian mentors and the productive socialization work they (could) do. In this tradition, success is based on building a well-designed interface that affords the right kind of capabilities. Third-wave HCI is more "critical" in its approach – seeking targeted design interventions that interrogate the ideological foundations undergirding practices. If the fundamental assumptions that gave rise to Wikipedia's social problems are not questioned, then a new software modification may only be a temporary fix to a larger systemic problem. While these two approaches are generally consider incommensurable, I'll present an example of a the design software technology that incorporates these two ways of approaching design in chapter 6.

## 1.5 CONCLUSION AND OVERVIEW

In this dissertation, I'll describe a body of work exploring individual decision making patterns and experimenting with software changes that change the environment in which Wikipedia editors operate.

In chapter 2, I'll describe a study of the decision making patterns that effect what type of changes are reverted in Wikipedia. I'll show evidence for behavioral patterns that reflect well documented social psychological phenomena (e.g. Ownership bias).

In chapter 3, I'll explore the effects of quality control actions on the work patterns of reverted editors and show how rejection within this review system predicts changes to the quality and quantity of a contributor's future participation. I'll also show evidence that the effects of rejection on new editors is substantially more pronounced than on more experienced editors.

In chapter 4, I'll build on of my analyses from chapter 3 to explore the system-level effect of rejecting newcomers contributions over time. I'll present a system-level, timeseries analysis which shows quantitative evidence that Wikipedia's shifting focus toward efficient quality control has caused the population of editors to abruptly shift from growth to decline.

In chapter 5, I'll explore a strategy for extending participation by adding a free-form feedback interface inspired by social learning theory. I'll show that the quality of contributions submitted via this extension is resilient to changes in where and how participation is elicited and explore the potential for such a system to boost Wikipedia's primary contribution (edits to articles) rates.

Finally, In chapter 6, I'll describe an intelligent user interface designed to support socialization practices with the intention of reversing trends observed in chapter 4. I'll present the results and lessons learned from a 3 month ethnographic design process, field study and evaluation. I'll provide recommendations for engineering software systems to support the work practices of a community and conclude with implications that this work has for software support for socialization in other open production systems and for the practice of HCI design.

Part II

CONTRIBUTIONS

## 2 HOW QUALITY, EXPERIENCE AND OWNERSHIP PREDICT REJECTION

---

### 2.1 INTRODUCTION

One of the key components of Wikipedia's success as an open production system is the review process through which contributions are rejected or accepted. While Wikipedia is generally not thought of as a peer review system since any contribution can be made and saved instantly, Stvilia et al. argued that the open editing system constitutes an informal peer review that moderates the quality of articles [82].

This process is informal and, to an outsider, appears disorganized. However in practice, the system appears to be robust and effective. In 2007, Preidhorsky et al. showed that 42% of damage is repaired before it's viewed even once and 70% of damage is repaired before it is viewed 10 times [74].

Traditionally, peer review has been used as a threshold for quality *before* publication. For instance, academic conferences in Computer Science typically have independent reviewers 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 is published while low quality work is rejected (or sent to lower quality venues). One important research question to ask is: How effective are these peer review processes at selecting for high quality submissions?

Previous work in evaluation of formal peer review systems either determined the quality of reviews by expert evaluation [39] or checking for significant differences in reviewer evaluations [89]. In either case, the evaluation focuses on a system's ability to select for high quality content despite bias.

The research covered in this chapter explores the effectiveness of the peer review system within Wikipedia by examining how the characteristics of editors and their work predict which contributions will be rejected. The goal this study is 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 study to the state of the art. First, I developed an automated measure (word persistence) for evaluating the quality of individual contributions. Using this automated measure, I examine whether words that have become established as high quality are difficult to change. I also test whether the recent quality of editors' work predicts whether their new work will be rejected.

Second, we'll look at the experience of an editor as a predictor of whether a contribution will be rejected. Decades of research show that individuals, groups and organizations all exhibit "learning by doing", where by their ability to perform complex tasks

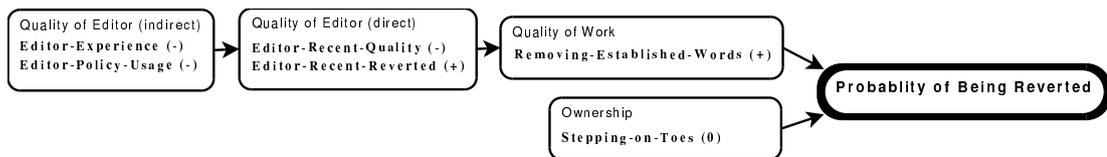


Figure 2: A model of how various factors should affect the probability of being reverted in an ideal system. (+) represents a positive correlation, (-) represents a negative correlation and (o) represents no correlation.

improves with their experience (see [4] for a review). I explored whether Wikipedia editors exhibit such a learning effect and quantitatively refute the premise that contributors produce more acceptable work as they gain experience.

Third, I'll show evidence that ownership bias has a strong, independent effect in Wikipedia. The number of *toes stepped on* by a contribution – i.e., the number of editors who would be likely to notice that an edit has removed a word that they had added – is a powerful predictor of whether that contribution will be rejected independent of the quality and experience of the editor making the contribution. Since the ownership of content is openly discouraged [94], this result demonstrates a non quality-related factor which has a strong effect on the outcome of review.

The rest of this chapter is organized as follows. First, I'll enumerate a set of six hypotheses about Wikipedian behavior in the context of related work. Next, I'll describe how measures of quality, experience, ownership and active editors are used in this analysis. Then, I'll present results and discussion for each of the six hypotheses and summarize the explanatory variables through a logistic regression model. Finally, I'll close with a summary.

## 2.2 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 **reversion** 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. A **reverts** represents the rejection of the intervening edits whose effects were removed.

I 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 2 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.

### 2.2.1 *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.

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

### 2.2.2 *Direct Editor Quality*

Previous research has explored what work is most valued by Wikipedia editors [6]. Other research has found that the structure of editor contributions affects the quality of articles [44]. 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 [74]. However, there is very little research that has directly explored the quality of an editor's individual contributions. In this study, I build on past measures of quality, value and reputation by developing an 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 [14]. I test whether a similar property holds true for Wikipedia.

**Hypothesis: 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. [45] and Brandes et al. [8] developed 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 actions should be interpreted [93]. 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 persistent 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.

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

### 2.2.3 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 [12]. Domains as diverse as automotive manufacture, pizza delivery and medical practice 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 [14].

If Wikipedians do become more effective editors as they gain experience, we should see a learning effect in Wikipedia in the form of a decreased probability of being reverted over time.

**Hypothesis: 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 received a lot of attention. Bryant et al. interviewed several of the most prolific editors to examine their motivations and growth [9]. They found that Wikipedians tend to transition from purely article construction to more organizational work, such as writing and discussing policies, as they gain experience. Kriplean et al. [48] and Beschastnikh et al. [6] explain and quantitatively present the use of policy and other internal mechanisms by editors to encourage, explain and discourage various community behavior. Similarly, in an analysis of talk page 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 discussion 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 [91].

In this analysis, I sought to determine the importance of an editor's command of Wikipedia policies and processes in determining whether their work was rejected. Assuming that editors who cite policies are at least familiar with them, they should be more likely to edit with the rules of Wikipedia.

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

#### 2.2.4 Ownership

Some user-generated content systems use the structure of ownership to make decisions about what changes will and will not be allowed. For example, open source software projects distribute 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 [61].

Fighting for controlling ownership of articles is openly discouraged in Wikipedia [94]<sup>1</sup>. Despite this, Kriplean et al.[48] and Thom-Santelli et al.[87] 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.

If editors feel ownership over the content that they add to articles, removing content that was added by a user who is likely to notice should increase the probability of being reverted. While feelings of ownership are natural and well studied [88], the strong presence of an ownership bias in Wikipedia's quality control process brings into question both the effectiveness of policy and the system cost of the content that is unnecessarily rejected.

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

## 2.3 METHODS

For my analysis, I used a random sample of approximately 1.4 million revisions attributed to registered editors (with bots<sup>2</sup> removed) as extracted from the January, 2008 database snapshot of the English version of Wikipedia made available by the Wikimedia Foundation<sup>3</sup>. In the results, I compare the analysis of 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, I combined them into a logistic regression with a boolean outcome variable representing whether a revision was eventually reverted<sup>4</sup>. Where I plot a probability of being reverted, I include a 95% confidence interval based on a normal approximation to the binomial distribution.

<sup>1</sup> <http://en.wikipedia.org/WP:Ownership>

<sup>2</sup> 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 I exclude them from my analysis.

<sup>3</sup> <http://download.wikimedia.org/enwiki/>

<sup>4</sup> Self-reverts, where an editor reverts himself, were not counted as reverts.

### 2.3.1 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, less than 25% of articles in Wikipedia were assessed a rating<sup>5</sup> and only 5% of articles had a rating higher than “start” [44], 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 [49]. However, Barnstars suffer from similar problems in that they are rarely given and often do not suggest which individual edits are being praised.

For my analysis, I required 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, I 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 [74]. I use a different metric: the number of editors who changed the article without removing the word. The number of revisions is preferable to the number of seconds because low quality words may survive many months without careful consideration in articles that are seldom revised. The number of revisions is preferable over the number of views under the assumption that a revision is more likely to be correlated with *critical reviews* of an article’s content. Given this assumption, the more reviews a contribution survives, the higher its quality. Therefore, my measure of lifespan is the number of revisions that a word survives.

To study a contribution across time, I measured 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 filtered out wikitext<sup>6</sup> and stopwords. I also filtered stopwords, but I’ve opted to include an article’s wikitext because substantial changes to an article can be made with wikitext alone. Further, wikitext is added and removed in the same way as visible content, therefore, it can be reviewed in the same way.

<sup>5</sup> <http://en.wikipedia.org/wiki/Wikipedia:1.0/A?oldid=255031288>

<sup>6</sup> 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.

Revisions	PWR
1: Apples are red.	2+4=6
2: Apples are blue.	0
3: Apples are red.	0
4: Apples are tasty and red.	1=1
5: Apples are tasty and blue.	0

Figure 3: **Persistent Word Revisions.** The word persistence values for five revisions of a sample article about apples are presented with arrows that show how words persist between revisions. Stop-words are greyed out since they are not considered in the algorithm. Revisions #3 reverts back to revision #1 and restores the word “red”.

Since a revert restores the state of an article, the 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 for two reasons: 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 total revert activity [45].

My mechanism for finding difference between revisions, the attribution of words to editors that add them and re-attributing words when reverts occur matches the methods used by Priedhorsky et al[74]. The key difference between their measure of the persistent word *view* and my measure of persistent word *revision* is that, rather than measuring the views that take place during the life of a word, I count the revisions in which the word continues to persist.

**Quality of an edit.** As a measure of the quality of an entire edit I 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, I 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 3 describes my 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, I 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, I controlled for the number of revisions left in an article by subsampling based on the amount of reviews possible after a current revision. I found no appreciable difference between the usefulness of PWR in my subsample and simply taking the log of the PWR across the full sample.

$$\text{PWRpW} = \frac{\sum_{\text{word}}^{\text{words}} \log \text{PWR}(\text{word})}{|\text{words}|} \quad (3)$$

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

**Verification.** To check my assumption that word persistence is an appropriate measure of quality, I 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. I performed a regression that mimicked the one performed by Kittur et al [44], with the addition of the scaled average word persistence. I 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<sup>th</sup> 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 my assumption that word persistence measures the quality an editor’s contributions.

### 2.3.2 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. I am 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.

I measured the number of interactions an editor has had in the system by grouping edits together into sessions. I 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<sup>7</sup>. I collapsed 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, I 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, I use only talk page citations in my model.

### 2.3.3 *Measuring ownership and active editors*

In order to test **HYP *Stepping on Toes***, I needed to determine how many active editors have their words removed by an edit. I 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 2.3.1, the word persistence measure tracks content through the history of an articles revisions. This allows us to associate editors with the individual words which they have added to an article several revisions later.

Determining the an editor’s status as “active” in an article is less straightforward. Since the watchlists of editors are not included in the database snapshot provided by Wikimedia, I 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, which is unlikely. By using recent editing 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.

## 2.4 RESULTS AND DISCUSSION

### 2.4.1 *Quality of work*

**HYP *Removing Established Words***. In order to measure how established a word has become, I 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, I used the word persistence algorithm described in section 2.3.1. Since this measure should be independent of the number of words changed during an edit, I included the number of words added and

<sup>7</sup> At the time of this study, the hour cutoff for sessions was purely intuitive. However, some of my more recent work provides quantitative reassurance for this intuition[29].

removed in the model (summarized in section 2.4.5) to control for large amounts of change.

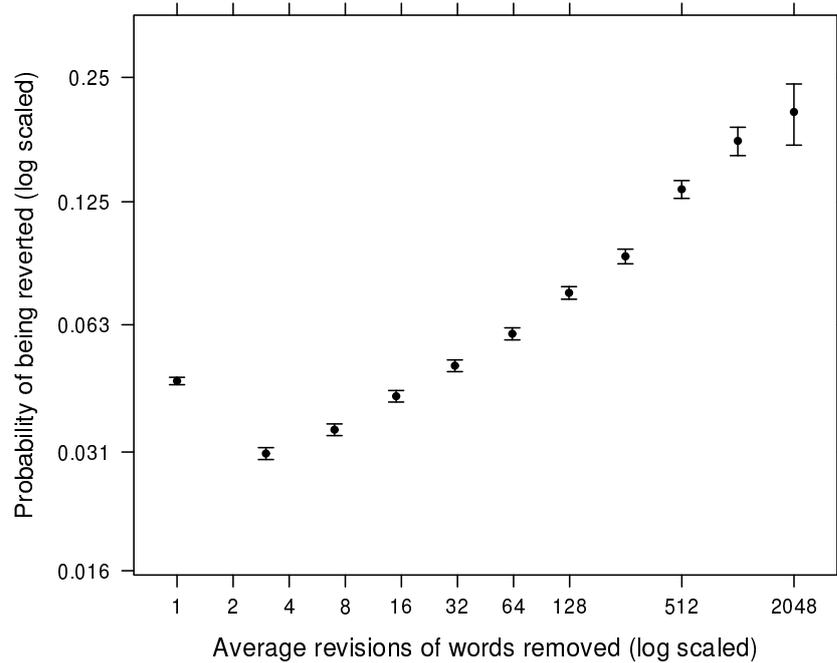


Figure 4: **Revert probability by removal of established words.** The probability of being reverted is plotted by the average persistence of words removed (as measured by word persistence).

Figure 4 shows that the probability of being reverted increases logarithmically as the average number of revisions that the removed words have survived increases logarithmically. Note the initial spike in the probability of being reverted for edits that remove words that were added in the last revisions. 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.

#### 2.4.2 Direct editor quality

**HYP Editor Recent Quality.** As a measure of the quality of a contribution, I use the word persistence metric described in Section 2.3.1, 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. I 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, I 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 flow of real-time.

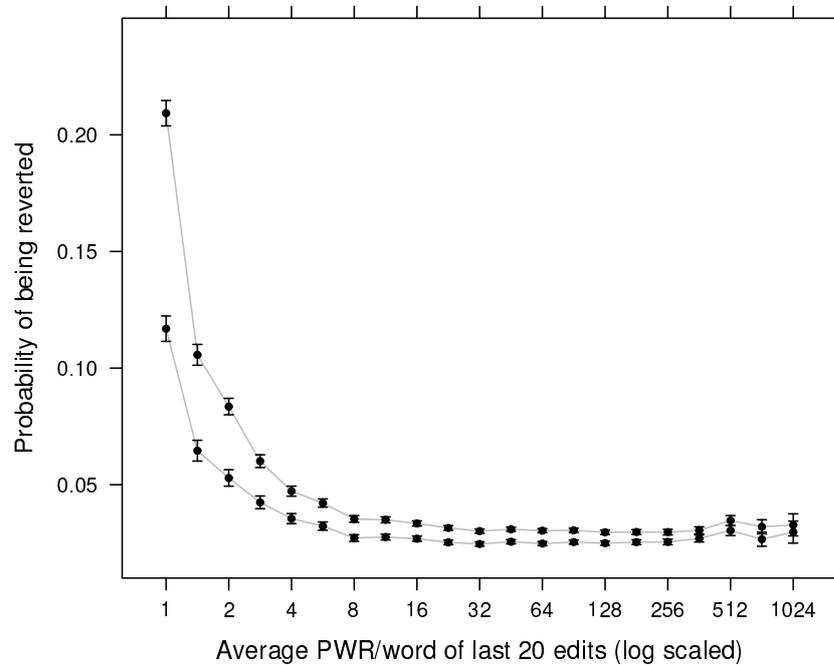


Figure 5: **Revert probability by recent word persistence.** The probability of being reverted is plotted by the average word persistence of the editor’s last 20 edits. The higher points represent the full sample while the lower were plotted after controlling for vandalism.

In order to ensure that the results were not simply an effect of vandalism, I required a way to control for the amount of reverts that were caused directly by vandalism. To perform this normalization, I examined reverting edit comments in order to detect which edits were reverted for vandalism<sup>8</sup> and scaled for the amount of vandalism that I was unable to detect based on numbers discovered by a manual coding performed by Priedhorsky et al. [74]. I make the assumption that vandalism that is not detected through edit comments is distributed similarly to vandalism that is detected. Figure 5 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 I ran the vandalism-controlled

<sup>8</sup> Vandalistic reverts were detected by looking for references to vandalism in the edit comments of the reverting revision with the D\_LOOSE algorithm introduced by [74].

subsample through the model, it confirmed that recent quality continues to be a strong predictor even when the effects of vandalism are removed.

**HYP Editor Recent Reverted.** This hypothesis is interesting because it seeks to answer something very basic about my 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.

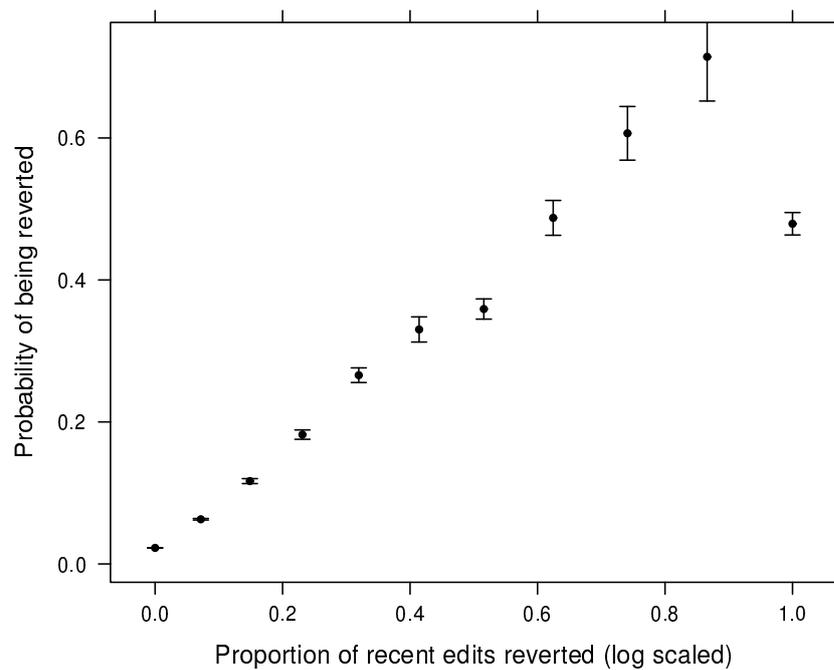


Figure 6: **Revert probability by recent reverts.** The probability of being reverted is plotted by the proportion of the editors last 20 revisions that were reverted.

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*, I decided to use the proportion of edits reverted over the last 20 edits performed by the editor as the explanatory variable. Figure 6 plots the proportions of recent revisions reverted by the probability that a subsequent revision will be reverted. As I 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. The 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 2.4.3). 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, I 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 the 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.

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

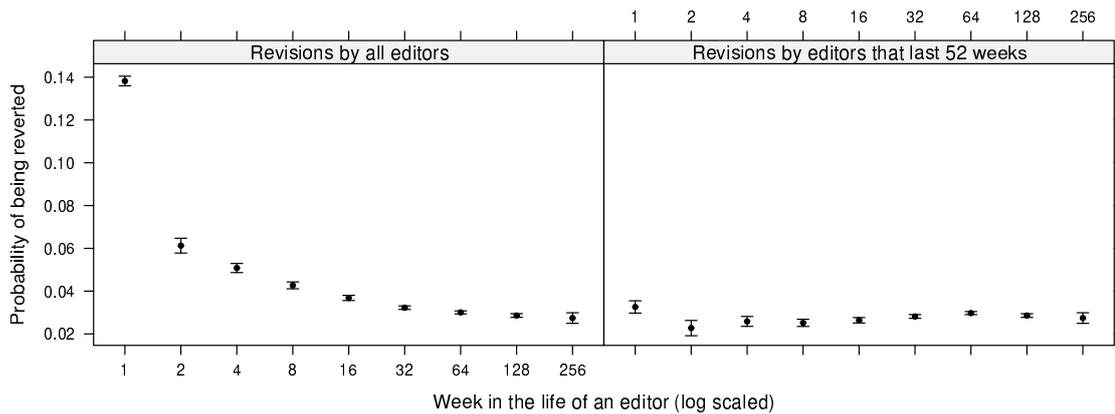


Figure 7: **Revert probability by editor tenure.** The probability of being reverted is plotted by weeks since editor started editing split into subsets based on how long the editors will eventually survive.

To determine whether I was seeing the effect of learning through experience within the system, I created a subsample of editors who last at least a year in Wikipedia. Figure 7 is a side-by-side plot of the complete sample and the subsample of editors who will last. In the full sample plot, the probability of being reverted falls when I sampled revisions by editors with more experience, but the subsample plot shows no appreciable fall in this probability through the life of editors. When I added the total number of days the editor would eventually be active to the model (see Table 1), tenure became an insignificant explanatory variable ( $p = .37$ ). I 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 these 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. The model, however, did detect a slight significant increase in the probability of being reverted with experience when I 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 [9]. 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, I 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, I used a simple regular expression that matched either “WP:<policy name>” or “Wikipedia:<policy name>”.

My 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. I performed this analysis under the assumption that only those editors with knowledge of policy would reference it. This 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.

#### 2.4.4 Ownership

**HYP Stepping on Toes.** In order to gather those editors who will notice when their words are removed, I used the active editors detection method described in section 2.3.3. This 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, I consulted the model and the 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 8 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 graphs that shows an expected probability as high as 0.5. In other words, depending

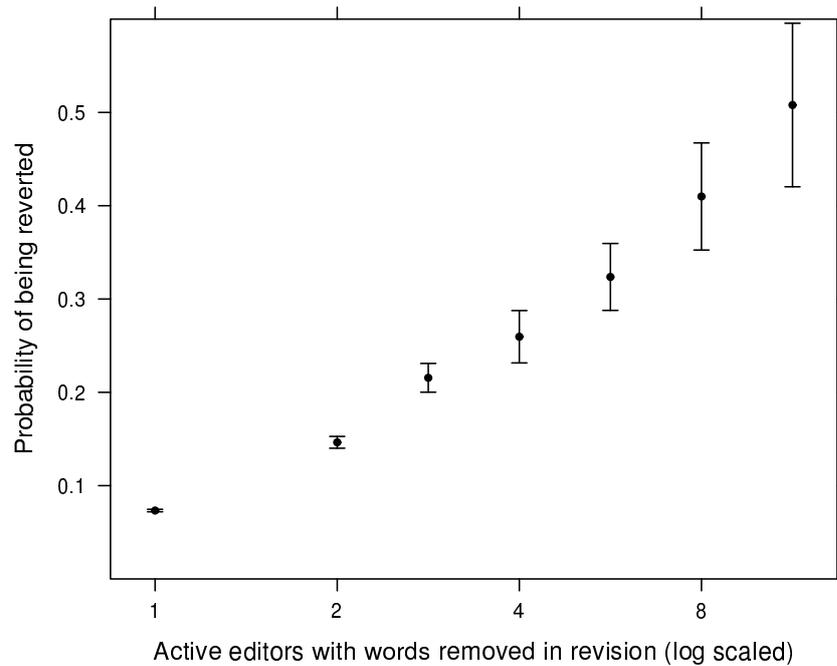


Figure 8: **Revert probability by toes stepped on.** The probability of being reverted is plotted by the number of active editors with words removed.

on the number of active editors whose words are removed by the current edit, the probability of being reverted can rise 50%.

The number of active editors with words removed was particularly interesting since, because it was the only feature used in the regression model that did not lose any power for any of the subsets I performed the regression over. 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 low quality, inexperienced editors who are reverted often. These results strongly confirm **HYP Stepping on Toes**.

#### 2.4.5 Grouped Analysis

In order to compare the effect of the metrics against each other to ensure their independence and significance I 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.

I used this model to answer questions about how the explanatory variables interact and to compare their power and utility for predicting when a particular revision will be reverted. I'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, I performed the regression over subsets that are not listed, but simply do not have space to show them here.

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	< .001	-3.640	.009	< .001
total days	-0.098	.010	< .001	-0.062	.011	< .001
word persistence	-0.183	.007	< .001	-0.161	.008	< .001
current tenure	-0.008	.009	.374	0.029	.010	.006
policy citations	-0.018	.012	.131	-0.024	.015	< .001
completed sessions	-0.252	.010	< .001	-0.159	.010	< .001
edits reverted	0.318	.004	< .001	0.276	.004	< .001
edits reverted (vand)	0.217	.003	< .001	0.128	.004	< .001
reverting edits	0.053	.004	< .001	0.049	.005	< .001
toes stepped on	0.231	.004	< .001	0.249	.005	< .001
est. of removed words	0.164	.006	< .001	0.128	.007	< .001
words added	0.024	.004	< .001	0.023	.004	< .001
words removed	0.006	.002	.008	0.005	.002	.048
w. persist.:est. of rem.	0.016	.005	.002	0.019	.006	.001

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<sup>9</sup>. 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 ( $r = .62$ ). I used the variance inflation factor of the model to determine that multicollinearity was low for all explanatory variables ( $< 2$ ).

Table 1 shows 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 *toes stepped on*. These changes suggest that as editors have been editing the Wikipedia system for a longer period of time, their history of being reverted, number

<sup>9</sup> Multicollinearity is a statistical phenomenon where two highly correlated variables cause erratic behavior in the individual estimates of a regression.

Table 2: Correlation table of explanatory variables.

	1	2	3	4	5	6	7	8	9	10	11	12
1. total days	1.0	.13	.58	.00	.40	-.13	-.12	.09	.02	-.06	-.01	.00
2. word persistence	.13	1.0	-.07	-.03	-.22	-.11	-.09	-.17	-.13	.10	-.08	-.02
3. current tenure	.58	-.07	1.0	.02	.62	-.12	-.16	.19	.06	.03	.00	.00
4. policy citations	.00	-.03	.02	1.0	.04	.00	.00	.03	.00	.01	.00	.00
5. completed sessions	.40	-.22	.62	.04	1.0	-.15	-.14	.32	.04	-.05	.02	.00
6. edits reverted	-.13	-.11	-.12	.00	-.15	1.0	.09	.03	.09	.10	.03	.02
7. edits reverted (vand)	-.12	-.09	-.16	.00	-.14	.09	1.0	-.01	.02	.05	.02	.01
8. reverting edits	.09	-.17	.19	.03	.32	.03	-.01	1.0	.13	.03	.02	.01
9. toes stepped on	.02	-.13	.06	.00	.04	.09	.02	.13	1.0	.13	.07	.07
10.est. of removed words	-.06	.10	.03	.01	-.05	.10	.05	.03	.13	1.0	.00	.01
11.words added	-.01	-.08	.00	.00	.02	.03	.02	.02	.07	.00	1.0	.03
12.words removed	.00	-.02	.00	.00	.00	.02	.01	.01	.07	.01	.03	1.0

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.

## 2.5 SUMMARY

In this study, I examined factors that seem likely to influence the probability that a contribution to Wikipedia will be rejected. Figure 2 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 2 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. I constructed a regression model that includes key measures of all four of these factors and controls for many of the likely confounds.

Table 3 shows the six hypotheses with a high level evaluation of the findings. I 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.

The results strongly support **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 [90]. I also found strong support for **HYP Stepping on Toes**, that the more active

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

Hypothesis	Support
<i>Removing Established Words</i>	Strong
<i>Editor Recent Quality</i>	Strong
<i>Editor Recent Reverted</i>	Strong
<i>Editor Experience</i>	Mixed
<i>Editor Policy Knowledge</i>	Weak
<i>Stepping on Toes</i>	Strong

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. I've implemented a proof of concept<sup>10</sup> using the "gadget" system in Wikipedia to demonstrate that such a UI change is computationally feasible, however a rigorous study of the system is not planned.

When evaluating **HYP Editor Recent Quality**, I found three pieces of evidence that support the assumption that word persistence (as measured by the *word persistence* metric) is, in fact, a useful approximation of the quality of an editor's contributions. First, 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 persistence editors. Second, the word persistence of an editor's recent work is a strong predictor of whether or not 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 chapter 3, I use this word persistence measure to detect changes in the overall productivity of editors.

**HYP Editor Recent Reverted** directly answers the question, "Is being reverted a quality of an editor?" My 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. In chapter 3, I show evidence that editors can and do adjust the rate at which they are reverted by making less *bold* edits after being reverted.

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

<sup>10</sup> <http://en.wikipedia.org/wiki/User:EpochFail/HAPPI>

I controlled for how long editors will continue to edit and how many sessions they will eventually complete. This change in predictive power occurs because of the high dropout rate of new editors who are reverted often. Because of this dropout effect, I judge the evidence for **HYP Editor Experience** to be mixed since I found no evidence of a learning effect in Wikipedia editors as they gain experience despite the usefulness of experience as an explanatory variable.

The hypothesis for which I 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 my 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, artificially inflating the number of reverts. A measure of controversy, such as those developed by [93], can be used to test whether this result holds when controlling for controversy.

## 3 HOW REJECTION AFFECTS EDITORS' WORK

---

### 3.1 INTRODUCTION

One of the keys to Wikipedia's success has been the ability to gather contributions from a large, diverse community of volunteer editors. To obtain the participation of such a wide array of web users, Wikipedia eases the transition from reader to editor via the technical platform and supporting policies; for example, the MediaWiki software lowers participation barriers by letting editors contribute through their web browsers; and community policies enable editing of most pages without even registering as a member.

However, this same ease of editing can be a double-edged sword, making it easy not only for legitimate editors to contribute but also for malicious or biased contributors to degrade the quality of existing content. Wikipedia addresses this issue through revert functionality. When an inappropriate edit is made to an article, any other editor can revert the article back to its previous state. The ability to easily revert changes alters the participation cost structure such that it costs less time/effort to fix a damaging edit than it costs to make the damaging edit in the first place.

Unfortunately, past research has shown that modifying the work of Wikipedia editors can reduce their rate of contribution in the future [102]. Reverts are a particularly direct form of modification, so it seems likely that reverts will have similar, if not more extreme, negative consequences on motivation. According to my analyses, the total percentage of reverts has increased over time to approximately 10% of all edits as of 2010. In this chapter, I'll focus on the effect that reverts, other than those labeled as vandalism or self-reverts, have on editor behavior. In just 2009, there were nearly 900,000 reverts of this type, an enormous amount of potentially lost effort and motivation.

#### 3.1.1 *Research questions*

The Collective Effort Model (CEM) provides a useful framework for characterizing the motivations of Wikipedia editors[41]. The CEM suggests a relationship between individual and group outcomes. For example, tasks that may produce high group value can increase individual motivation; furthermore, increasing group outcomes can increase individual motivation for other tasks by making the individual feel that overall group effort is producing more value.

Getting reverted may make individuals feel that their contributions are not valued by the group and not leading to positive group outcomes, resulting in demotivating effects. In this case, being reverted might cause editors to produce less work after the

revert.

**RQ1: How does being reverted affect the quantity of editor work?**

One perspective that has not been examined much in the literature is that being reverted could be part of the learning process for editors. When an editor is reverted, they may reconsider the edits they make thus improving their work. Wikipedia's guideline for editing boldly<sup>1</sup> encourages editors to make changes as they see fit and let the collaborative process help them check the quality of their work. Therefore, we might expect to see that editors who are reverted learn something from the experience and produce higher quality work afterward.

**RQ2: How does being reverted affect the quality of editor work?**

One way that Wikipedia editors might avoid motivation loss when they are reverted is by discussing the situation. Commiseration with other editors might help preserve or even reinforce editors' motivations to continue participating in Wikipedia. Wikipedia's Bold, Revert, Discuss cycle<sup>2</sup> encourages editors to engage in discussion with other editors after their work has been reverted. However, if they are too strongly discouraged by having been reverted they may communicate less, losing the opportunity to learn and improve.

**RQ3: How does being reverted affect communication?**

Wikipedia's population of volunteer editors appears to have abruptly entered a steady decline in recent years. The source of this decline appears to be reduced newcomer retention [84, 99]. For example, while nearly 40% of new editors remained active for a year pre-2005, that number dropped to only 12-15% post-2007 [99]. A significant challenge to recruiting and retaining newcomers is the difficulty they experience in understanding the vast history of prior contributions, decisions, policies, and standards that the community has evolved over time. Such factors have led to newcomers being reverted at higher rates than more experienced editors, a trend that has been getting exacerbated over time [84]. Could it be that reverts of newcomers are responsible for the reduction in active contributors?

The experience of editors who perform reverts is also an interesting variable for analysis too. Editors who have been working in Wikipedia for a long time will have more experience interacting with other editors. This experience could give them insight into how to discard other editors' work without demotivating them. More experience could also make it more difficult to relate to newbies, especially in the changed context of the more crowded, more complete Wikipedia of today.

**RQ4: How does experience moderate the effects of reverts on contribution?**

<sup>1</sup> <http://en.wikipedia.org/wiki/WP:BOLD>

<sup>2</sup> <http://en.wikipedia.org/wiki/WP:BRD>

### 3.1.2 *Structure and contributions*

In this chapter, I'll present an analysis of the impact that reverts have on editor behavior. Although prior studies have examined reverts, this is the first study to quantify the impact of reverts on editor behavior. Understanding these effects is crucial to understanding how emerging editorial policy in Wikipedia is creating a high-quality encyclopedia.

This study considers 400,000 revisions and uses variety of techniques to analyze the performance of those editors. Lest the contributions get buried in the details, I summarize them here. My most important findings are that (1) reverts do have a negative impact on editor contribution and survival, especially for newcomers; and (2) when editors do continue to contribute after a revert, the quality of their contributions increases. Taken together, these findings suggest a more nuanced view of reverts that both recognizes the benefits of reverts for learning while acknowledging their costs in editor motivation and contribution, especially for newcomers. This view has practical implications for system design as well as for the design of intelligent tools for supporting reverts that enhance the potential for learning while minimizing the potential for demotivation.

## 3.2 RELATED WORK

A number of studies have looked specifically at reverts in Wikipedia. One common subject of research has been quantifying revert activity. This has often focused on reverting "damaged" content, such as the effects of vandalism. Wikipedia editors have developed sophisticated semi-automated processes for detecting, escalating, and sanctioning vandalism [30], with the result that vandalism is typically reverted very quickly, often on the order of minutes [45, 58, 74, 90]. The prevalence of reverts has been growing over time [45], as has a trend towards reverting newcomers [84].

Although vandalism is an important challenge, reverts often signify more substantive disagreements. Significant prior research has also focused on detecting, visualizing, and understanding the editing dynamics that involve reverts and their relationship to conflict. Conflict is not unexpected in a user-generated knowledge base, in which hundreds of thousands of contributors each bring their own viewpoints and knowledge, often leading to factions, and territoriality<sup>3</sup> [45, 87]. Tools to visualize conflicts and viewpoint differences based on reverts or revisions have also been developed [8, 19, 45, 83, 90].

Relatedly, there have also been a number of studies developing explicit trust and quality metrics that leverage the longevity of an editors' contributions (e.g., [17] and [82]). For example, Adler and de Alfaro[1] derive author reputation from the survival of an author's edits over time and surface this information through an interface that color-codes the content they contribute. Zeng et al. [101] use dynamic Bayesian net-

---

<sup>3</sup> See also, my results from chapter 2.

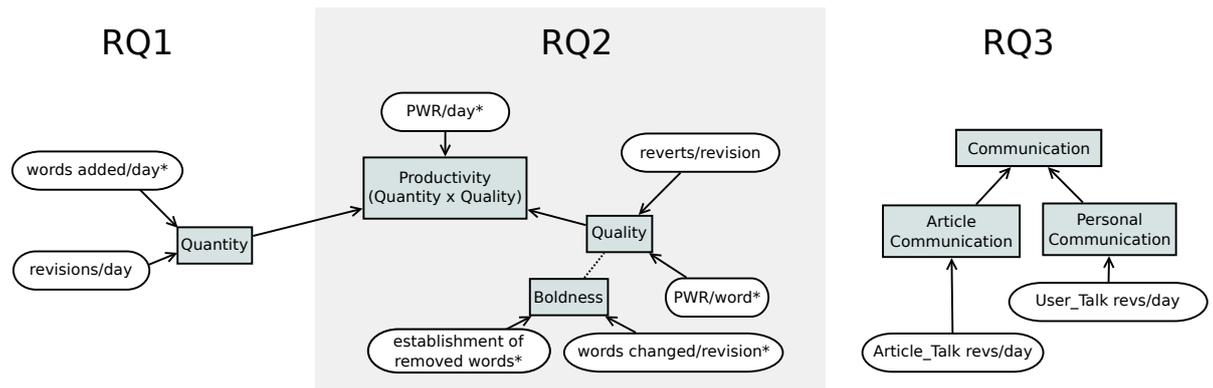


Figure 9: Hidden variables of editor activity are connected to the metrics that were used as proxies in the analysis and are divided by the research questions they are used to explore. Metrics obtained from 2008 data are signified with an “\*”. An arrow from A to B means “A is used as a proxy for B”. The dotted line between boldness and quality represents the confounding effect described in section 3.3.4. RQ4 does not rely on hidden variables.

works to calculate the evolution of trust in an article using the editing status of authors and inserted/deleted text as input. Explicit visualization of revert activity in a page’s history (among other features) has also been shown to decrease readers’ perceptions of the trustworthiness of a page [44].

In this study, I’ll contribute to this foundation of research by quantifying the impact of reverts on the behavior of the reverted editor. Although much is known about revert activity at a macro level, little is known about its impact on editor behavior. Some studies have looked at the effect of other kinds of feedback on contribution. Zhang & Zhu [102] conducted an analysis of editors who created new articles in Wikipedia, finding that editors are less likely to contribute when another editor edits the page they created, irrespective of whether they were adding or removing content. This effect was mitigated by the amount of experience the article had. Choi et al. [13] examined the effect of socialization tactics on newcomers to WikiProjects<sup>4</sup>, finding that negative (constructive) feedback but not positive feedback led to relative increases in subsequent contribution. These studies suggest that interacting with an editors’ contributions – either directly or through explicit feedback – can have an impact on the editor’s subsequent activity. However, neither Zhang & Zhu nor Choi et al. examine the impact of reverts, nor changes in the quality or boldness of contributions in addition to their quantity.

### 3.3 METHODS

#### 3.3.1 Dataset

To look for the effects of being reverted, I gathered a sample of 400,000 revisions saved by registered editors from the January, 2010 data dump of the English Wikipedia by combining two subsamples. The *reverted* subsample comprises 200,000 revisions that were reverted by another editor (no self-reverts) and not tagged as vandalism<sup>5</sup>. The *not-reverted* subsample comprises 200,000 revisions that were not reverted to be used as a control for comparison. These samples were gathered in such a way that no two revisions were performed by the same editor. This restriction ensures that the effects observed are not dominated by a few prolific editors. For the rest of this paper, I refer to the reverted and not-reverted revisions in the combined sample as *sampled edits*.

I used the word persistence metadata (described below) generated for the analysis described in chapter 2 and intersected it with the *sampled edits* to obtain a second dataset with 176,438 revisions (90,641 reverted, 85,797 not-reverted). Figure 9 notes which metrics are based on this smaller dataset.

#### 3.3.2 Quantity

Editors perform work in Wikipedia by editing pages. Starting an article, contributing to an existing article, sending a message to another editor and asking a question are all represented within the system as an edit that creates a new revision of a page. I'm primarily interested in the quantity of work as it applies to the construction of Wikipedia's encyclopedia articles. I measure the quantity of article work in two ways: revisions/day and words added/day.

*Revisions/day* is the number of article revisions created by an editor per day. This metric quantifies work based on the rate than an editor makes changes to articles.

*Words added/day* is the number of non stop-words added to articles by an editor per day. This metric quantifies work based on the rate than an editor adds content to articles.

#### 3.3.3 Quality

I approximate the quality of an editor's contributions based on how other editors react to it. If Wikipedia's natural review mechanism of collaborative editing is effective in selecting for quality content (see [82] for more explanation), higher quality contributions should be more likely to be accepted by other editors in Wikipedia. I define two metrics for the quality of editor contributions: reverts/revision and PWR/word.

<sup>4</sup> WikiProjects are subproject groups within Wikipedia that focus on a narrow content areas or activity types. For example, WikiProject Fungus is focused on improving articles about mushrooms, molds and other fungi.

<sup>5</sup> Reverts for vandalism were detected using the D\_L00SE approach from [74].

*Reverts/revision* is the proportion of an editor's revisions that have been reverted in a given timespan. Reverting a revision is an indication that the reverting editor does not consider the revision to be of acceptable quality.

*PWR/word*, based on the word persistence metric described in chapter 2 (see figure 3 in chapter 2 for an example application of PWR), is the average number of revisions that words added by an editor persist. Higher quality contributions should add words that, on average, survive more reviews by other editors.

#### 3.3.4 *Boldness*

Both of the metrics used as proxies for the quality of an editor's work are measures of whether other editors revert their contributions. Under these metrics, an editor can appear to do work of higher quality by making "safer" edits that are less likely to be reverted by other editors. Since the "boldness" of an edit can change the interpretation of results related to the quality metrics, I looked for changes in boldness in response to reverts. To quantify the boldness of contributions, I measured two characteristics of edits: the number of words changed and the average establishment of words removed.

*Words changed per revision* is the number of not stop-words added or removed by an edit. Edits that change more article content are considered to be more bold than edits that change little or no content.

*Establishment of words removed* is the average PWR of words that an edit removes. Words with high PWR have survived many revisions, and I assume, the scrutiny of other editors who edit the article. The higher the PWR value, the more strongly the content has become an established part of an article, and the bolder an edit that removes it.

#### 3.3.5 *Productivity*

In addition to measuring how many words a user adds to Wikipedia ("Quantity") and how long those words last on average ("Quality"), I'm also interested in finding a way to combine these metrics to estimate how much of an impact a user is having on the encyclopedia. Such a measure should include the amount of output the editor produces as well as the quality of such output as was discussed in equation 1 presented in section 1.4.1.

On average, an editor who adds more words that last longer is affecting Wikipedia more than an editor who adds fewer words, or whose words are removed more quickly. I therefore define a metric for the *productivity* of an editor as:

*PWR/day*: the product of the number of not stop-words words added to articles and the number of subsequent revisions those words persist.

Of course, any such measure is an approximation of actual impact, since, at the extremes, an editor who focuses on improving Wikipedia policy might add significant value without ever contributing a single non-stop-word to an article. However, over a

large number of edits, and a large and diverse set of editors, it is likely that measuring changes in how many words are contributed and in how long they last will serve as an effective proxy for productivity.

### 3.3.6 *Measuring communication*

For this analysis, I am interested in two types of communication: communication about article content and personal communication between editors. Wikipedia makes a clear distinction between these types of communication via the namespace<sup>6</sup> in which the communication occurs. Communication about article content occurs as edits to pages in the Article\_talk namespace and personal communication occurs as edits to pages in the User\_talk namespace.

*Article\_talk revisions/day* and *User\_talk revisions/day* are the number of revisions to pages in the Article\_talk and User\_talk namespaces created by an editor per day. Changes to the revisions/day in these namespaces should reflect changes in the amount of an editor's communication.

### 3.3.7 *Measuring changes*

To identify changes in the activities and characteristics of reverted editors, I directly compare measures of the activities of the editors before and after a sampled edit. To establish a pre-revert state, I analyze the edits an editor made in the week preceding the revert. I then compute the change in activity from this baseline level for each of the following four weeks of that editor's activity. I refer the difference in each week the *activity delta* ( $\Delta$ ). I chose a week as the time frame for analysis due to the observation that many editors have weekly periodic edit profiles, perhaps in response to work and other "real life" activities.

When measuring activities that vary widely from day to day for editors, I divided the activity delta by the standard deviation of the activity/day in the pre-revert week to produce a *controlled activity delta* ( $\Delta/\sigma$ ). This ensures that all editors (regardless of overall amount of activity) are expressed in the results equally.

### 3.3.8 *Notes on figures and tables*

The plots in this chapter include standard error bars. For many of the plots, the error bars may be too small to see due to the high number of observations sampled – but they are present. I used regressions many times throughout this study. The most important regressions are summarized in table 4, and discussed throughout section 3.4. The results of other regressions are presented with coefficients and significance values where appropriate, but the regression tables are redacted to save space.

<sup>6</sup> To differentiate between the different types of edit activity, Wikipedia has several "namespaces" devoted to different types of content.

## 3.4 RESULTS AND DISCUSSION

## 3.4.1 RQ1: How does being reverted affect the quantity of editor work?

As discussed in section 3.1, theory predicts that being reverted could demotivate editors, leading them to do less work in the future. To measure changes to the quantity of work of an editor, I use revisions/day, the average number of revisions an editor completed per day. To check the validity of this metric, I also performed the same analysis over words added to articles per day, and found no significant difference in the apparent effects.

To identify the effect of being reverted on the quantity of work performed by reverted editors, I performed a linear regression over controlled article activity delta<sup>7</sup>. The “Activity  $\Delta/\sigma$ ” column of Table 4 reports that being reverted is a significant, negative predictor ( $\beta = -.292, p < .001$ ) of controlled article activity delta. I also performed a regression over a controlled activity delta for words added per day  $\Delta/\sigma$  and found that being reverted predicted a similar significant, negative effect ( $\beta = -.934, p < .001$ ).

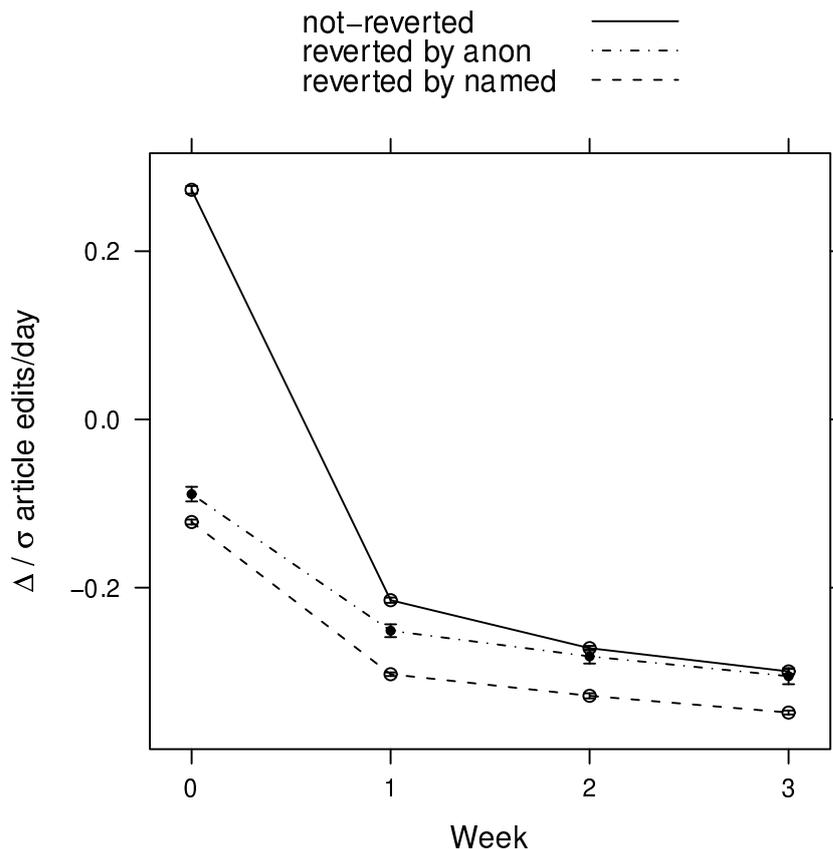


Figure 10: **Article activity  $\Delta/\sigma$ .** For four weeks after a sampled edit, the change in article activity is reported. Reverted edits are split by whether the reverting editor was registered or anonymous. A control group of similar editors who were **not** reverted is included for comparison.

<sup>7</sup> To control for other factors that could confound the analysis, I also included editor tenure and characteristics of the reverted edit as independent variables in the regression.

Table 4: Regressions over article activity  $\Delta/\sigma$ , survival and PWR/day  $\Delta/\sigma$  for four weeks after the sampled edits. Characteristics of the sampled edit's change to an article (words added, words removed, establishment of removed words) and whether it was a reverting edit itself or reverted back to by another editor are included in the regression to control for effects they could have on future work. For the discussion, statistical significance corresponds to  $\alpha = .025$ . Multicollinearity was checked for using correlation between explanatory variables. All correlation coefficients are below 0.5. PWR  $\Delta/\sigma$  is scaled and logged to normalize it.

	Activity $\Delta/\sigma$			Reverted activity $\Delta/\sigma$			Survival (logistic)			log(PWR/day $\Delta/\sigma$ )		
	$\beta$	SE	$P(>  z )$	$\beta$	SE	$P(>  t )$	$\beta$	SE	$P(>  t )$	$\beta$	SE	$P(>  t )$
(Intercept)	.149	.004	< .001	-1.149	.003	< .001	1.206	.006	< .001	.322	.013	< .001
Reverted(True)	-.292	.005	< .001				-.681	.008	< .001	.627	.021	< .001
Week since revert	-.178	.002	< .001							.021	.008	.013
Reverted editor tenure	.024	.002	< .001				1.594	.010	< .001	-.004	.010	.660
Words added	.062	.012	< .001				.004	.021	.844	.022	.019	.244
Words removed	-.003	.003	.309				-.017	.005	< .001	-.008	.009	.405
Est. of removed words	-.001	.003	.708				-.109	.006	< .001	.058	.014	< .001
Reverting(True)	.030	.012	.015				.326	.027	< .001	.399	.050	< .001
Reverted to(True)	-.037	.011	.001				.019	.021	.363	.022	.046	.633
Rvtd:Week since revert	.103	.003	< .001	-.074	.003	< .001				-.001	.012	.953
Rvtd:Reverted tenure	.043	.003	< .001	.064	.002	< .001	.373	.014	< .001	-.037	.014	.011
Rvtd:Reverting tenure				-.009	.002	< .001						
Rvtd:Words added	-.063	.012	< .001	-.000	.002	.872	-.047	.023	.041	-.035	.020	.084
Rvtd:Words removed	.002	.003	.746	-.002	.002	.253	.015	.006	.008	.031	.014	.029
Rvtd:Estab. of rm words	.003	.004	.367	.002	.002	.145	.045	.007	< .001	-.005	.016	.740
Rvtd:Reverting	.002	.014	.884	.026	.005	< .001	.334	.029	< .001	-.678	.057	< .001
Rvtd:Reverted to	.073	.013	< .001	.034	.006	< .001	.049	.025	.054	-.177	.057	.002
Rvtd:Reverting is anon				.012	.006	.032						

Table 5: Dependent variable characteristics. n for all dependent variables is 684,508 after removing non-finite values.

	Reverted	Week	Rvtd tenure weeks	Words added	Words removed	Estab. of rm	Reverting	Reverted to
$\mu$	0.50	1.00	21.01	41.28	73.70	30.95	0.09	0.07
$\sigma$	0.50	1.41	33.47	1082.78	2728.89	123.10	0.30	0.26
$\eta$	0.50	1.00	3.80	3.00	0.00	0.00	0.00	0.00

Figure 10 shows the controlled revisions/day delta of reverted editors split by whether the *reverting* editor was registered or anonymous beside the controlled revisions/day delta of not-reverted editors. In the first week after being reverted (week 0), not-reverted editors increase their activity about 0.3 standard deviations. By contrast, reverted editors decrease their activity by 0.1 standard deviations. By the second week after being reverted (week 1), the difference between reverted and not-reverted editor activity has decreased substantially. Although the average activity level of editors reverted by anonymous editors appears to converge by week 2, the activity of editors reverted by named editors does not converge in the four observed weeks. Essentially, editors reverted by anonymous editors recover to the same average level of activity within a couple of weeks, but those reverted by named editors do not recover for at least one month (if ever).

Since the plot shows that some editors do not recover in the month of time observed. I suspected that the long-term effect on reverted editors could have been due to editors being demotivated enough to leave Wikipedia entirely. To determine whether being reverted predicted a decreased survival rate in Wikipedia, I defined a simple metric for survival as a boolean variable that is True when an editor continues editing at least 8 weeks after the sampled edit and performed a logistic regression over it. The “Survival (logistic)” column of Table 4 reports that being reverted is a significant, negative predictor of survival ( $\beta = -.681, p < .001$ ).

The regressions over article activity and survival have shown us that being reverted predicts both a decrease in activity and a reduction in the probability of survival. Can the reduction in activity be explained by the decreased percentage of surviving editors or do editors who will continue to edit decrease their activity as well? To answer this question, I performed another regression over the controlled article activity delta for only the surviving editors. Being reverted remained a significant, negative predictor ( $\beta = -0.237, p < .001$ ). It appears that, even when editors will continue editing for at least two months, being reverted has a significant effect on the quantity of work they perform.

To measure the amount of the difference between the reverted editors in Figure 10 that is independently due to the anonymous or named status of the reverting editor, I performed a regression. This regression included only the reverted sampled edits because those are the only ones with a reverting editor. Column “Reverted activity  $\Delta/\sigma$ ” in Table 4 shows that the reverting editor’s anonymous status predicts a marginally significant, positive effect on future contributions ( $\beta = .012, p = .032$ ). There are two possible causes for this effect: a higher rate of contribution or a higher rate of survival among editors reverted by anonymous editors. To measure the independent effect of each of these two possible causes, I performed two additional regressions. I found that being reverted by an anonymous editor did not predict a higher rate of contribution for surviving editors ( $\beta = 0.011, p = .273$ ), but it did predict a higher rate of survival (logistic:  $\beta = .153, p < .001$ ). This result suggests that the apparent demotivating effect of being reverted is significantly less severe when the reverting editor is anonymous

and that the effect is largely due to an increased probability of survival.

Taken together, the results of this subsection suggest that being reverted both decreases the probability an editor will continue editing and decreases the motivation of those editors who do continue editing – at least temporarily.

The lesser demotivating effect in the case that the reverting editor is anonymous suggests that characteristics of the reverting editor could have an effect on the severity of the demotivation. Anonymous editors lack a persistent identity in Wikipedia and are likely to be perceived as outsiders with only a passing interest. In Wikipedia, allowing anonymous editors to contribute is controversial<sup>8</sup> and discrimination against anonymous editors has warranted community concern<sup>9</sup>. A possible explanation of the results is that reverted editors take the feedback of having their work discarded by anonymous editors less seriously.

#### 3.4.2 RQ2: How does being reverted affect the quality of editor work?

Though editors reduce the quantity of their work in editing articles after being reverted, it is possible that reverts serve as an opportunity for feedback and learning. Here we explore the question of whether getting reverted tends to increase the quality of an editor's future work.

**Quality.** As described in section 3.3.3, I first quantify the quality of editors contributions by measuring how likely they are to get reverted in the future. To ensure that the revert itself did not bias the results, that edit is dropped from the recent activity when establishing the pre-revert state for reverted editors. To check the validity of these results, I also performed this analysis over PWR/word and found no substantial difference in the apparent effects.

Figure 11 shows that editors who are reverted are less likely to get reverted in the future. Reverted editors immediately drop significantly in their probability of being reverted. The percentage drop of about 4% shown in the figure is substantial since it represents more than a third of the underlying revert probability (11%). By contrast, not-reverted editors become more and more likely to be reverted each week.

It is interesting to note that a week after the sampled edit, both reverted and not-reverted editors make edits that are increasingly likely to be reverted each week of continued editing – and at a similar rate. One interpretation is that the being reverted is a learning experience for the reverted editor and acceptability of their work to other editors after being reverted represents an increase in quality. Another interpretation is that editors decrease their boldness in order to make their work more acceptable after being reverted, since edits that are less bold (eg. copy-edits and other minor changes) should be less likely to be rejected by other editors.

<sup>8</sup> [http://en.wikipedia.org/wiki/WP:Editors\\_should\\_be\\_logged-in\\_users](http://en.wikipedia.org/wiki/WP:Editors_should_be_logged-in_users)

<sup>9</sup> [http://en.wikipedia.org/wiki/WP:IPs\\_are\\_human\\_too](http://en.wikipedia.org/wiki/WP:IPs_are_human_too)

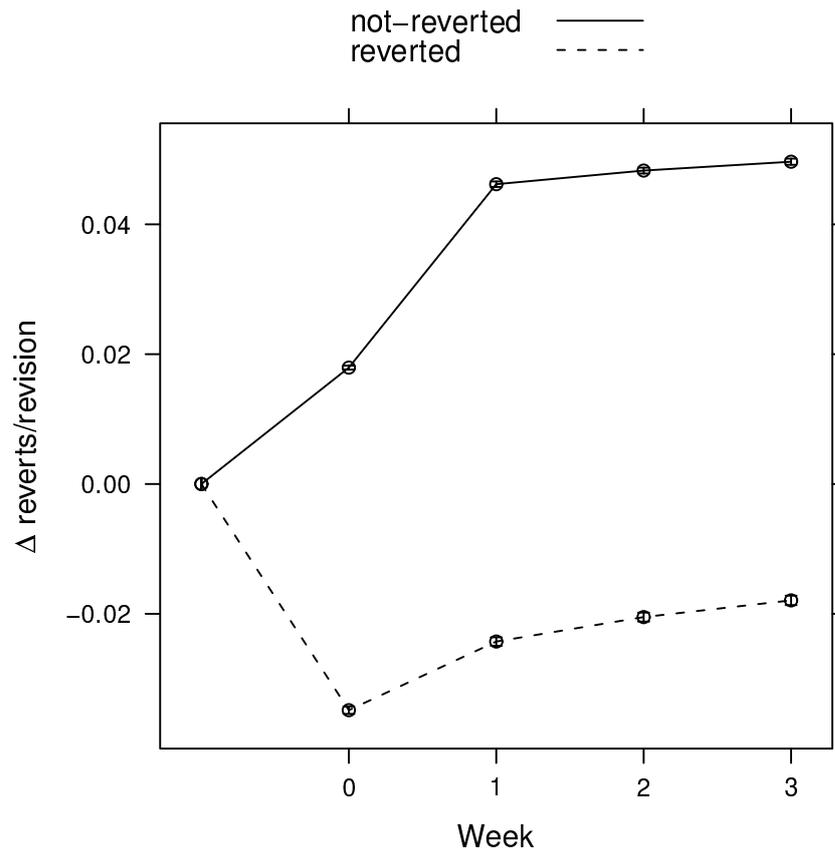


Figure 11: **Reverts per revision  $\Delta$** . For the four weeks immediately after a sampled edit, the change in quality of work as reverts/revision is reported. A control group of similar editors who were **not** reverted is included for comparison.

**Boldness.** To find out whether editors change the boldness of their contributions to articles, I measured two characteristics of the edits they made: words changed/revision and the establishment of removed words.

Figure 12a shows that reverted editors will change fewer words per revision and remove less established words after being reverted. The number of words changed per revision delta for reverted editors is consistently 75 – 100 below the delta for not-reverted editors. Interestingly, the difference in the establishment of removed words only diverges a week after being reverted. These results suggest that being reverted could be a check on the boldness of the type of edits being made and reverted editors could be improving their revert percentage by playing it safe.

**Productivity.** Does the change in boldness explain the change in reverts/revision or are editors also increasing the quality of their work? In order to determine whether editors only become less bold after being reverted or whether they increase in quality of their work, I measure how their productivity changes based on PWR/day, a measure of both the quantity and quality their work described in section 3.3.5.

Figure 13 plots the controlled PWR/day delta for reverted and not-reverted editors. Reverted editors increase in their productivity more than not-reverted editors immediately after being reverted and the difference stays relatively consistent through the four observed weeks. The “ $\log(\text{PWR}/\text{day } \Delta/\sigma)$ ” column in Table 4 confirms that being

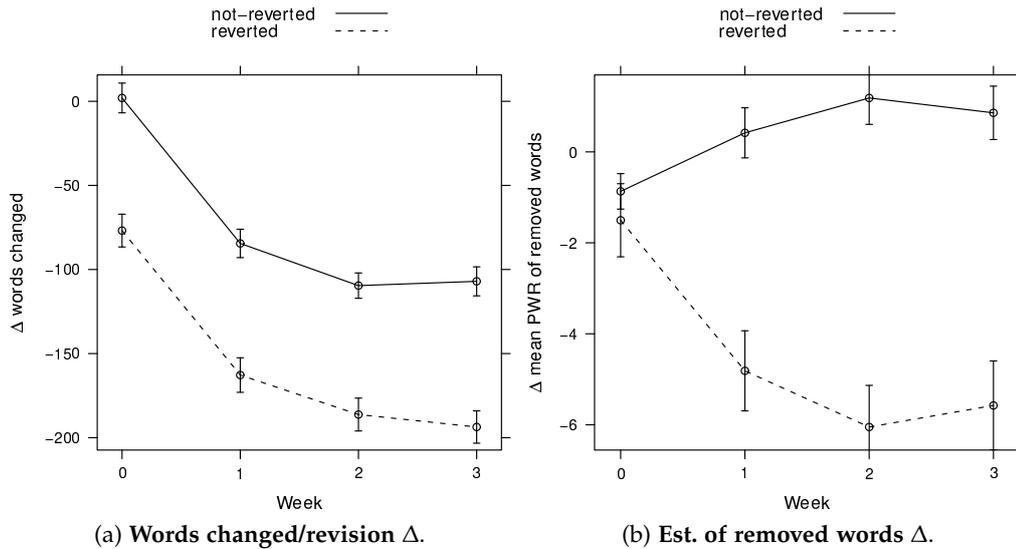


Figure 12: **Changes in boldness.** For the four weeks immediately after a sampled edit, the change in the boldness of work is reported via two metrics: words changed/revision and establishment of removed words. A control group of similar editors who were **not** reverted is included for comparison.

reverted is a significant, positive predictor of future productivity ( $\beta = .627, p < .001$ ) and the interaction between being reverted and week has a small, insignificant effect ( $\beta = -.001, p = .953$ ). The pre-revert PWR/day (not shown in Table 4 for formatting reasons) was included in the regression as a control for how productive an editor was before being reverted. The interaction between being reverted and recent productivity was a significant, negative predictor ( $\beta = -.412, p < .001$ ). This effect suggests that editors who are producing high PWR/day before being reverted do not gain as much PWR/day after being reverted.

Although the results reported in this section show that reverted editors were more likely to leave Wikipedia and that the ones who stay will become less active and less bold in their work, this result shows that they will increase their quality, and therefore their productivity enough to make up for the difference. However, highly productive editors do not benefit in the same way that less productive editors do from being reverted. This result suggests that the learning effect from being reverted is primarily experienced by editors who are not already very productive to begin with.

### 3.4.3 RQ3: How does being reverted affect communication?

When editors reduce their effort toward editing Wikipedia articles, do they move that effort toward other Wikipedia activities, such as communication? To answer this question, I applied the same approach that I used in section 3.4.1 for article activity to produce variance controlled activity deltas ( $\Delta/\sigma$ ) for Article\_talk and User\_talk. I limited my analysis to surviving editors since editors who do not continue work on Wikipedia are predisposed to not make edits to Article\_talk and User\_talk.

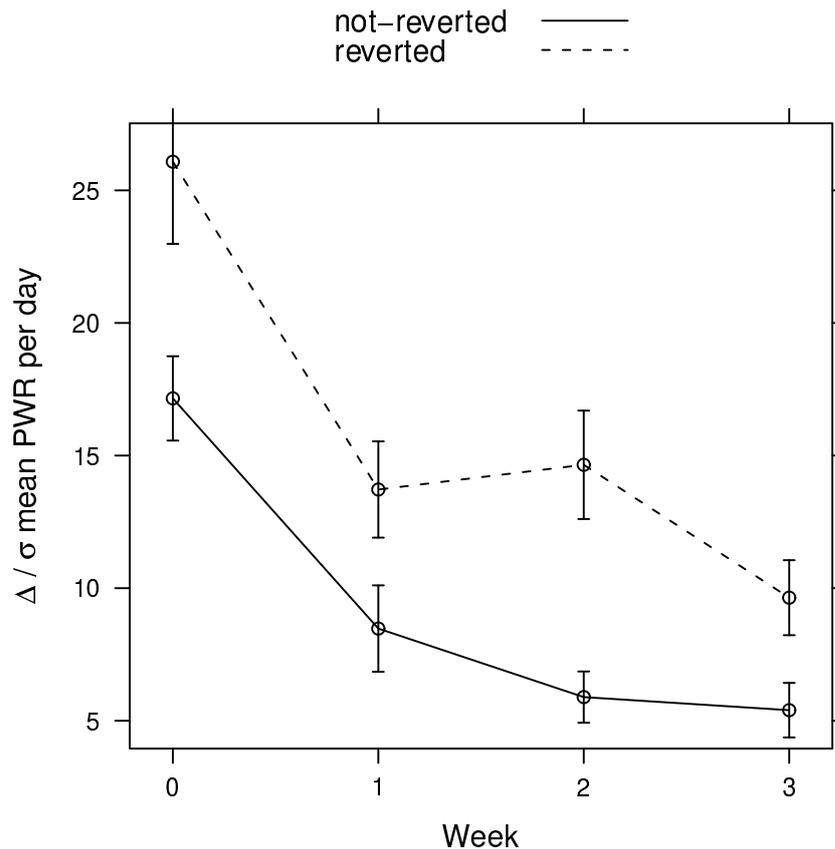


Figure 13: **PWR/day**  $\Delta/\sigma$ . For the four weeks immediately after a sampled edit, the change in productivity is reported as the controlled PWR/day delta. A control group of similar editors who were **not** reverted is included for comparison.

To determine whether being reverted was a significant, independent predictor of future communication, I performed two regressions over controlled activity deltas for of `Article_talk` and `User_talk`  $\Delta/\sigma$ . The regressions report a significant negative effect of being reverted on `User_talk` activity ( $\beta = -.079, p < .001$ ), but an insignificant effect of being reverted on `Article_talk` activity ( $\beta = -0.023, p = 0.663$ ).

It's interesting to note that reverted editor tenure was a powerful predictor of the change in `Article_talk` activity ( $\beta = .123, p < .001$ ). Since, reverted editor tenure is significantly different between the reverted and not-reverted subsamples (T-test:  $\text{diff}=76.29, p < .001$ ), I controlled the not-reverted sample to be distributed similarly by tenure<sup>10</sup> to the reverted subsample before plotting values for Figure 14.

Figure 14 plots the controlled activity deltas for surviving editors after controlling for reverted editor tenure. For `Article_talk`, the difference between reverted and not-reverted editors appears insignificant, but for `User_talk`, the difference is substantial (0.1 standard deviations) and reverted editors did not recover to the communication activity levels of not-reverted editors in the four weeks observed.

Overall we have seen that when editors were reverted they reduced their personal communications, but did not reduce their communication over article content significantly. For Wikipedia, this result may be good news, because it suggests that the

<sup>10</sup> To control for tenure differences, I matched proportions in quantiles between the two subsamples.

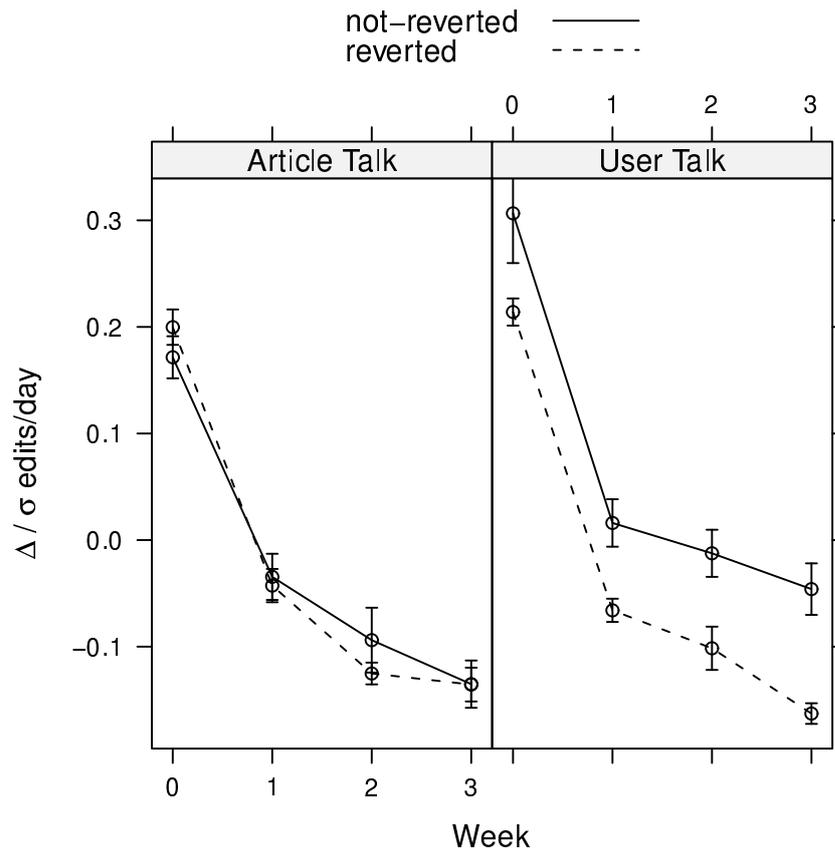


Figure 14: **Communication activity**  $\Delta/\sigma$ . For the four weeks immediately after a sampled edit made by surviving editors (defined in Section 3.4.1), the change in Article\_talk and User\_talk communication activity is reported. A control group of editors with similarly distributed tenure who were **not** reverted is included for comparison.

reverted editors are not being entirely demotivated from discussing article content in Wikipedia.

#### 3.4.4 RQ4: How does experience moderate the effects of reverts on contribution?

We now know that although being reverted reduces an editor's contributions, it also increases the quality of their work. To understand how the effects of being reverted are moderated by experience, I repeated the analysis from sections 3.4.1 and 3.4.2 in the context of the tenure<sup>11</sup> of both reverted and reverting editors.

##### 3.4.4.1 Reverted editor tenure

**Activity and survival.** Figure 15 represents activity changes for two interesting subsets of editors: *newbies* are editors with less than one month of tenure and *old-timers* are editors with more than one year of tenure.

After a revert, old-timer editors do experience a temporary reduction to their article activity, but they return to the level of activity of their not-reverted counterparts within

<sup>11</sup> The amount of time since an editor's first edit.

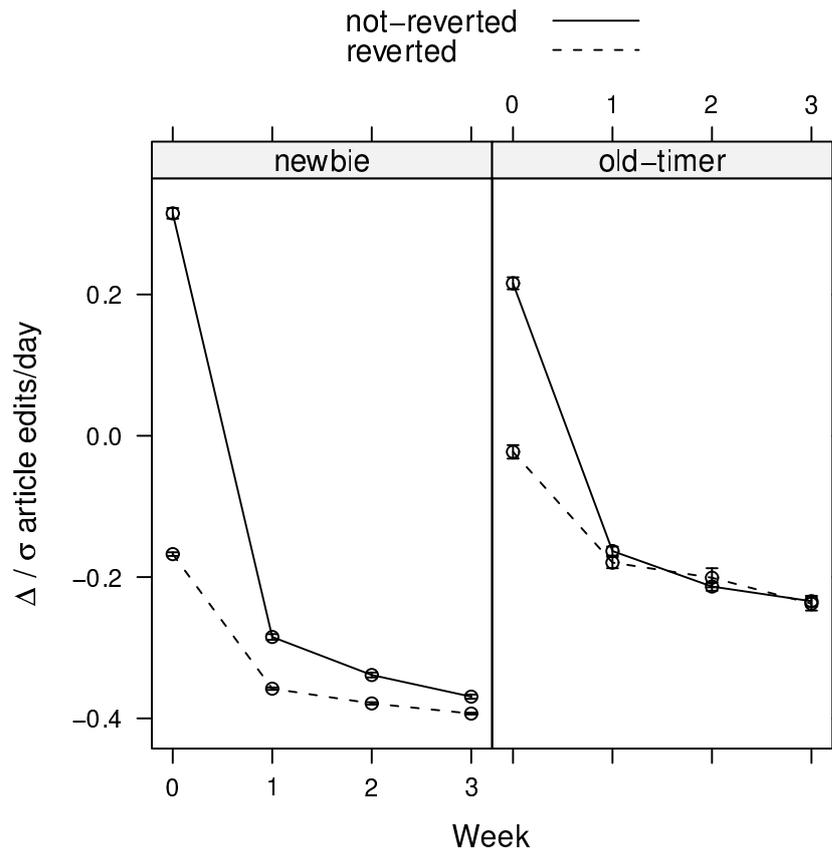


Figure 15: **Article activity  $\Delta/\sigma$  by reverted editor tenure.** For the four weeks immediately after a sampled edit, the change in article activity is reported for newbies and old-timers. A control group of similar editors who were **not** reverted is included in each graph for comparison.

two weeks of being reverted. For newbie editors, the difference in the activity delta is both stronger and longer-lasting. Reverted newbies take more than four weeks to return to the activity levels of not-reverted newbies.

The “Activity  $\Delta/\sigma$ ” column of Table 4 reports a significant, positive effect of reverted editor tenure ( $\beta = .043, p < .001$ ) on the controlled activity delta when an editor is reverted. This positive effect suggests that the more experience an editor has in the system the less their article activity will decline after being reverted. The “Survival (logistic)” column in Table 4 also reports that reverted editor tenure is a significant, positive predictor of editor survival ( $\beta = .373, p < .001$ ). In summary, newbies are less likely to continue editing after being reverted than old-timers and the ones who do continue to edit reduce their activity more than old-timers.

**Quality and productivity.** The “log(PWR per day  $\Delta/\sigma$ )” column of Table 4 shows a significant, negative effect on controlled PWR/day delta for reverted editor tenure ( $\beta = -.037, p = .011$ ). That is, the more experience reverted editors have in Wikipedia, the lower the productivity boost they see after being reverted.

This result is surprising given the *Activity and survival* analysis above, that more experienced editors see less of a drop in contribution after being reverted than less experienced editors. A plausible explanation is that more experienced editors increase

the quality of their article contributions less. A linear regression over PWR/word delta, finds a marginally significant, negative effect of reverted editor tenure ( $\beta = -1.145, p = .180$ ), confirming that the smaller quality increase for editors with more experience can be explained by the lesser productivity boost they experience.

#### 3.4.4.2 Reverting editor tenure

**Activity and survival.** Figure 16 compares the controlled activity deltas of editors reverted by newbie and old-timer *reverting* editors. Here, we see the opposite effect of reverted editor tenure. While the activity of editors reverted by newbies will recover within two weeks, the activity of editors reverted by old-timers did not recover in the four observed weeks.

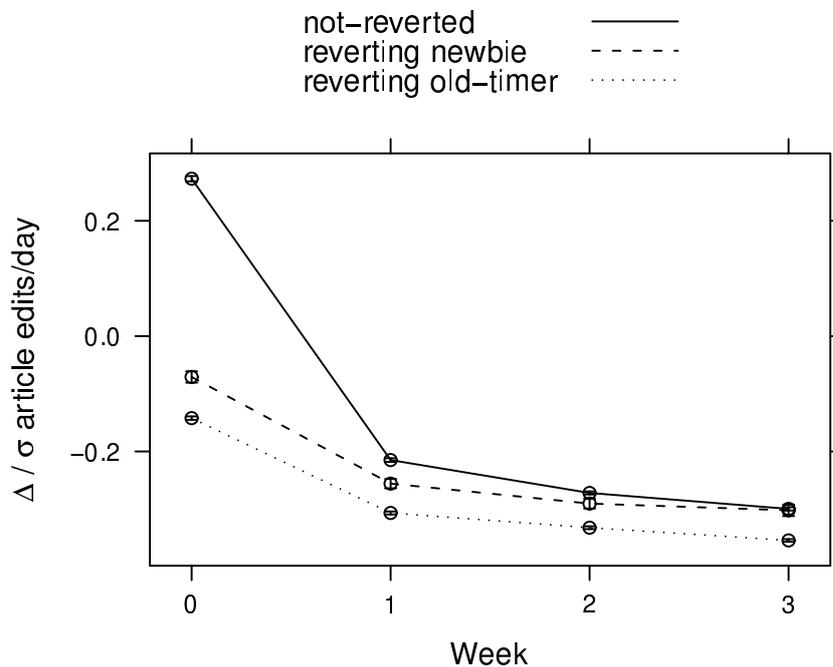


Figure 16: **Article activity  $\Delta/\sigma$  by reverting editor tenure.** For the four weeks immediately after a sampled edit, the change in article activity is reported. Reverted edits are split by whether the reverting editor was a newbie or old-timer. A control group of similar editors who were **not** reverted is included for comparison.

The “Reverted activity  $\Delta/\sigma$ ” column in Table 4 shows a significant, negative effect for reverting editor tenure on the controlled activity delta ( $\beta = -.009, p < .001$ ). A logistic regression over the survival of reverted editors also showed a significant, negative effect for reverting editor tenure ( $\beta = -.075, p < .001$ ). This suggests that the longer the reverting editor has been participating in Wikipedia, the lower the likelihood that the reverted editor will continue to make edits. Even for editors who will survive, more reverting editor experience means a larger reduction in the quantity of work the reverted editor will perform after being reverted.

This result suggests that editors with a high level of experience are not as effective in maintaining the motivation of the editors they revert than editors who are less familiar

with wiki-work. Another interpretation is that old-timers have an enhanced ability to identify unconstructive editors and chase them away.

**Quality and productivity.** Linear regressions over my metrics for quality and productivity of reverted editors did not report significance for the tenure of the reverting editor. This suggests that quality and productivity changes in the reverted editor are not significantly affected by the experience of the reverting editor.

### 3.5 CONCLUSIONS

The goal this research was to understand how reverts affect the motivation of reverted editors. Since it is difficult to measure motivation directly, I instead measure indirect effects that might be caused by changes in motivation, including reducing the amount of contribution, or the ultimate reduction, withdrawal from Wikipedia.

I found that editors are more likely to withdraw from Wikipedia after being reverted and that the ones who stay decrease their quantity of work. These effects are especially strong for reverted newbies and when the reverting editor has more experience in Wikipedia. This finding is especially relevant for Wikipedia since it has a well documented problem with decreasing newcomer retention [84, 99].

On the other hand, editors who continue to do work in Wikipedia after being reverted appear to increase the quality of their work. This effect is especially true for newbies and less productive editors. Being reverted appears to be a learning experience, helping the editors who need it most learn to be more effective Wikipedians. Under my measure of productivity (PWR/day), the net effect of reverts on Wikipedia is positive: on average, an editor who is reverted produces more persistent words per day – even including those editors who withdraw from Wikipedia in the calculation!

Editors also change their communication patterns after being reverted. Although there was an overall decrease in the communication activities of reverted editors, editors do not appear to decrease their communication about article content after being reverted. The sustained article communication activity is a positive sign. Wikipedia's Bold, Revert, Discuss cycle encourages article discussion as a reaction to being reverted, and discussion should help reverted editors learn how to improve their work. However, editors did decrease their communications to other editors after being reverted. This reduction may be a sign of withdrawal risk: personal communications with others in a community can reinforce incentives to participate and reinforcement may be exactly what is needed for editors whose contributions were rejected.

Taken together, these results suggest that overall reverting activity in Wikipedia is healthy and valuable, with the training effects dominating the demotivating effects. However, there are specific cases in which reverting activity might be managed better, to dampen the negative effects and amplify the positive. I offer the following recommendations to designers of open production communities like Wikipedia:

**Support rebuffed users.** I found that being reverted often precedes a reduction in

participation. Furthermore, the reverted editors decrease their communication with other editors at a time when they are vulnerable to leaving the community. Perhaps having other users reach out to them could help reinforce their connections to the community. Support should be personal since the results of other research suggests that impersonal socialization tactics can do more harm than help [13].

**Encourage the learning effect.** I found that being reverted predicts an increase to the quality and productivity of an editor's article work. The negative feedback appears to be an opportunity for users to improve the quality of their participation. Viewing feedback like reverts as an opportunity to teach should be both encouraged and supported. While performing a revert, the reverting editor should be encouraged to provide clear feedback to help the reverted editor improve.

**Focus on newcomers.** I found that newcomers are particularly likely to decrease their contributions after they are reverted. I also saw some evidence that they can learn the most from being reverted. Newcomers should be reached out to actively to help them become socialized into Wikipedia.

### 3.5.1 *The impact of these metrics for measuring wiki-work*

It's worthwhile at this point to note that since the publication of these results, nearly all of the metrics and analysis strategies developed for this research project have become part of the standard set of metrics used by the Wikimedia Foundation to measure editor activity in Wikipedia<sup>12</sup>.

---

<sup>12</sup> See <http://meta.wikimedia.org/wiki/Research:Metrics>

## 4 THE RISE AND DECLINE OF WIKIPEDIA

### 4.1 INTRODUCTION

Open collaboration systems like Wikipedia require a large pool of volunteer contributors. Without volunteers to power the subsystems, Wikipedia itself would cease to function. Like any volunteer community, open production systems need to maintain an inner circle of highly invested contributors to manage and direct the group. However, with statistical predictability, all contributors to such systems will eventually stop contributing [100, 72].

The success of an open projects appears to be highly correlated with the number of participants they maintain. Projects that fail to recruit and retain new contributors tend to die quickly [18]. In order to maintain a pool of contributors, newcomers must be continually socialized into the organization. Some of these newcomers must move from the periphery of the community to the center [9, 73].

Historically, Wikipedia has managed this process effectively. The community grew from hundreds of active editors in 2001 to thousands in 2004 and peaked in March of 2007 at 56k active editors (see figure 17). Suh et al. describe this growth as a self-reinforcing mechanism: as Wikipedia became more valuable, the project attracted more contributors to increase its value [84].

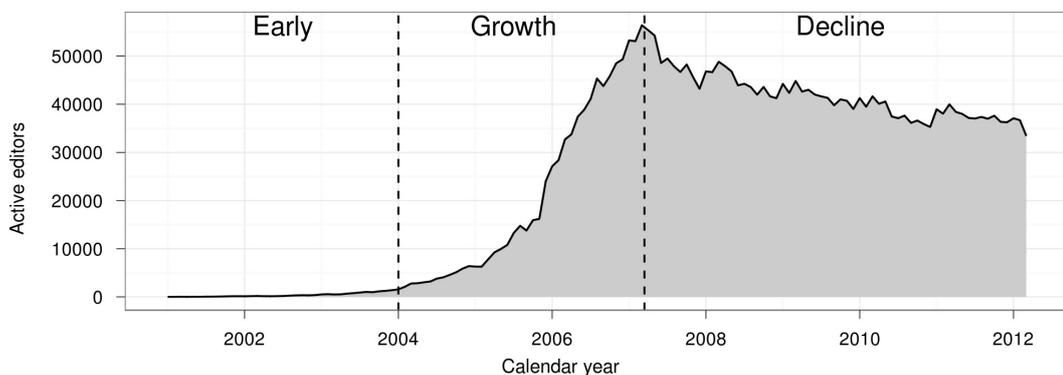


Figure 17: **The English Wikipedia editor decline.** The number of active editors ( $\geq 5$  edits/-month) is plotted over time for the English language Wikipedia.

Then, at the beginning of 2007, things changed. Participation entered a period of decline [99]. Why? Recent research suggests different explanations. Suh et al. argue that the decline could be the result of the depletion of available work in the context of a population model [84]. However, of Wikipedia's "Core 1000" most important articles

are still of poor quality, and across the encyclopedia, only 14,072 (0.362%) articles are rated “good” quality <sup>1</sup>, so there’s certainly still work to do.

Other researchers point to failed socialization systems. Indeed, evidence suggests that it is difficult for newcomers to find work to do [47] and to discover where to ask for help<sup>2</sup>. Generic, standardized socialization tactics (such as generic welcome messages) are common on Wikipedia, but these tactics are demonstrably less effective at encouraging sustained contribution than personalized variants [13]. Wikipedians have organized project groups to support socialization, but they fail to serve most newcomers [66].

Another possibility is that the volunteer community that powers Wikipedia could simply be “right-sizing”. Perhaps now that the most important work of the encyclopedia is done, there is no need for the 56k editors who were active in 2007. Two pieces of data argue against this theory. First, as noted above, the vast majority of articles in Wikipedia are still below community standards for “good” articles. Second, under-represented groups still find it challenging to join. For instance, one study found that only 9% of edits are made by female editors, and that articles of particular interest to women are shorter than articles of interest to men [52]. Until editors are representative of the population of potential contributors, it is difficult to argue that the socialization practices are sufficiently effective.

In this chapter, I’ll define a type of Wikipedia editor that I’ll call a *desirable newcomer* (see section 4.3.4). The first few edits of these newcomers indicate that they are trying to contribute productively (i.e. acting in good-faith) and, therefore, are likely to become valuable contributors if they continue editing. I’ll show that, while the proportion of desirable newcomers who arrive at Wikipedia has been holding steady in recent years, a decreasing fraction of these newcomers continue editing past their first *session*. I’ll then demonstrate that this declining retention of desirable newcomers has been caused, at least in part, by the Wikipedia community’s reactions to the enormous influx of contributors between 2004 and 2007.

During Wikipedia’s growth period, the community’s views toward the goals of the project changed. These new views were instantiated in a set of policies, and a suite of algorithmic tools were developed for enforcement. Over time, these changes resulted in a new Wikipedia, in which newcomers are greeted (rudely) by automated quality control systems and are overwhelmed by the complexity of the rule system. Since these changes occurred, newcomers – including the crucial, desirable newcomers – have been leaving Wikipedia in droves.

This paper makes three contributions to understanding the declining retention in this context. First, I’ll implicate Wikipedia’s primary quality control mechanism [82], the rejection of unwanted contributions, as a strong, negative predictor of the retention of high quality newcomers and show that these newcomers’ contributions are being rejected at an increasing rate. Next, I’ll show how algorithmic tools, which

<sup>1</sup> <http://enwp.org/WP:FAS> - <http://enwp.org/WP:GAS>

<sup>2</sup> [http://meta.wikimedia.org/wiki/Research:New\\_user\\_help\\_requests](http://meta.wikimedia.org/wiki/Research:New_user_help_requests)

were built to make the work of controlling the quality of Wikipedia’s content more efficient, exacerbate the effect of rejection on desirable newcomer retention and circumvent Wikipedia’s conflict resolution process. Finally, I’ll show how calcification has made Wikipedia’s policy environment less adaptable and increased the difficulty of contributing to the community’s rules – especially for newcomers.

## 4.2 MOTIVATION AND HYPOTHESES

### 4.2.1 *Rejection of newcomers*

Stvilia et al. argues that Wikipedia’s open contribution system constitutes an informal peer review where all contributions are initially accepted; other editors perform review and reject unwanted contributions [82]. This review system is apparently effective at producing value.

Yet in chapter 3, I showed that this kind of rejection significantly reduces newcomers’ contribution rates. When considering this potentially demotivational effect of reverts in the context of increased rejection for newcomers observed by Suh, et al. [84], it is tempting to conclude that rejection of contributions is scaring away newcomers. However, Halfaker et al. didn’t look for temporal effects, and although they controlled for vandalism reverts, they don’t control for the quality of the contributors and thus could not draw conclusions about the necessity of the revert itself.

Thus, these observations could be explained by a decline in the quality of newcomers. Such a decline could be caused by an early adopter affect, where users who were most interested in Wikipedia’s success flocked to the site when it was young. Perhaps later users are less devoted, and less likely to contribute productively. If such an effect were taking place, the rise in rejection of newcomer contributions would be a sign of health for the community. In other words, these observations could simply be the product of the Wikipedia’s review system doing its job.

However, there are many reasons to believe that the rate of rejection of newcomers’ contributions would increase regardless of changes in quality and intentions of newcomers. Suh et al. argues that the rising rate of reverts among all editors (including newcomers) could be attributed to increasing conflict over the amount of available work which naturally decreases as the encyclopedia reaches completion [84]. Indeed, the analysis in chapter 2 showed that editors were more likely to get into conflict when editing the same parts of articles.

Changes in the community’s views toward the project’s goals could also be a cause of increased rejection. For example, the definition of “unwanted” contribution has changed over time. While presenting at Wikimania in 2006, Jimmy Wales urged Wikipedians to change their focus from quantity to quality. His presentation signified a shift from Wikipedia as a catch-all for encyclopedia-like content to a more restrictive project. In a study of the birth and death rate of articles in Wikipedia, Lam et

al. observed that the rate at which new articles were rejected substantially increased following Mr. Wales's keynote [50].

There are also external pressures for Wikipedia to tighten its review process. After high profile cases of libel (e.g. the Seigenthaler libel incident<sup>3</sup>), the community strengthened norms and enforcement surrounding biographies of living persons. The official policy page states: "Contentious material about living persons that is unsourced or poorly sourced [...] should be removed immediately and without waiting for discussion." Since Wikipedia has historically benefited from an abundance of contribution, rejecting a few good contributions in favor of removing damage was seen as a reasonable trade off.

Over time, the encyclopedia may also be becoming more difficult to contribute to due to the increasing completeness of articles. An analysis I performed for the Wikimedia Foundation<sup>4</sup>, shows that recent newcomers are more likely to contribute to longer, more complete articles (4x longer in 2009 than 2004) and the length of the article at the time of contribution was a significant predictor of rejection.

I suspect that the increased rates of rejection are explained by changes in the way that Wikipedia deals with new content and that this pattern of rejection negatively affects the retention of desirable newcomers

**Hypothesis: Rejection & retention:** *Increasing rates of rejection have caused a decrease in the retention of desirable newcomers.*

As an examination of this hypothesis, I'll report new results that demonstrate the following:

- The quality of newcomers has not decreased substantially since the middle of Wikipedia's exponential growth.
- During exponential growth, the rate of rejection for edits made by desirable newcomers rose and the survival rate of desirable newcomers fell.
- Rejection of desirable newcomer contributions is a significant, negative predictor of retention.

#### 4.2.2 *De-personalized welcoming of newcomers*

The Wikipedia community has a long history of building algorithmic tools that operate on Wikipedia's content to serve a wide variety of needs. These tools can generally be divided into two categories: *robots* or *bots* are autonomous computer programs that perform edits with little or no human intervention; *human-computation* tools are extensions or standalone programs that enhance a user's ability to interact with the wiki platform, but still rely on human judgment to perform operations.

**Bots.** The roles of bots in Wikipedia have grown substantially in both size and scope since the early days of Wikipedia. The first bots enabled power users to perform many

<sup>3</sup> [http://en.wikipedia.org/wiki/Wikipedia\\_bibliography\\_controversy](http://en.wikipedia.org/wiki/Wikipedia_bibliography_controversy)

<sup>4</sup> [http://meta.wikimedia.org/wiki/Research:Newbie\\_reverts\\_and\\_article\\_length](http://meta.wikimedia.org/wiki/Research:Newbie_reverts_and_article_length)

repetitive activities faster than any human could manually. In 2006, Wikipedia administrator Tawker initiated a new genre: the vandal fighter bot. In order to deal with a coordinated attack by deviant users adding references to “Squidward” – a cartoon character – across the encyclopedia, Tawker built a bot that monitored and identified damaging changes to the encyclopedia in real-time using a simple text pattern matcher. This form of fast-paced content curation was quickly expanded to other easily-identifiable acts of vandalism. By 2012, counter-vandalism bots are in wide use. ClueBot NG, currently Wikipedia’s most prolific counter-vandalism bot, uses machine learning and neural network approaches to identify and reject over 40,000 acts of vandalism a month, with a median time to revert of five seconds. However, despite the use of state-of-the-art techniques, only the most egregious vandalism can be caught by these fully autonomous workers

**Human-computation tools.** To efficiently catch the damage that bots miss, a number of tools were developed to more efficiently re-introduce human judgment into the vandal fighting task. Some tools, like Twinkle and rollback, extend the basic functionality of Wikipedia’s web-based interface, adding contextually-relevant buttons and links that automate tasks for a human user. For example, from an article’s revision history, an editor with Twinkle installed can remove all of another editor’s most recent contributions to an article and send a pre-written message telling the editor not to vandalize the encyclopedia again. Standalone tools, like Huggle, organize a well defined set of tasks into one interface, such as the presentation of suspected vandalism edit “diffs”<sup>5</sup> and the ability to approve or reject edits with a single click.

These algorithmic tools have apparently made quality control both more efficient and more effective. Previous work has shown that the duration during which vandalism is visible in an article has been decreasing [45, 74]. These tools also reduce the amount of volunteer effort that must be devoted to rejecting unwanted contributions by organizing work into a queue and performing several algorithmic operations for each human operation.

However, recent work suggests that the efficiency of these tools may have some negative impact on the experiences of a newcomer. An analysis performed by Geiger found that newcomers generally find their newly-created articles are deleted faster than they can contribute to them<sup>6</sup>. A related study that I performed with Geiger showed that these algorithmic tools have been taking an increasing role in “welcoming” newcomers via warning messages [32]. By late 2007, over half of new users received their first message from an algorithmic tool. That figure grew to 75% by mid 2008.

Although the use of algorithmic tools appears to have dramatically increased the efficiency of Wikipedia’s quality control system, I suspect that the use of these tools to reject contributions has been negatively affecting the retention rate of desirable newcomers due to their impersonal nature and the aggressive rejection they encourage.

<sup>5</sup> “diff” refers to the visual presentation of the changes made by a single edit to an article.

<sup>6</sup> [http://meta.wikimedia.org/wiki/Research:The\\_Speed\\_of\\_Speedy\\_Deletions](http://meta.wikimedia.org/wiki/Research:The_Speed_of_Speedy_Deletions)

**Hypothesis: Tool use & consequences:** *The use of algorithmic tools to reject newcomer contributions is exacerbating the decrease in desirable newcomer retention.*

As an examination of this hypothesis, I'll report new results that demonstrate the following:

- The use of algorithmic tools to reject newcomer contributions has been increasing.
- The use of algorithmic tools by old-timers to reject the contributions of newcomers correlates strongly with a breakdown in Wikipedia's preferred conflict resolution process.
- The use of algorithmic tools to revert newcomer edits significantly increases the negative effect of rejection on desirable newcomer retention.

#### 4.2.3 *Calcification of norms against newcomers*

Research conducted during Wikipedia's growth period has drawn links between Wikipedia's success and editors' ability to participate in the creation, modification and enforcement of the rules that govern editing. As the editor community grew implicit norms were formalized into a growing corpus of official rules and procedures [11], and rule creation and enforcement became increasingly decentralized [6, 25].

The trends towards decentralization and norm formalization in Wikipedia governance may have been a natural and healthy responses to community growth [25]. Formally documenting community practices facilitated wider dissemination in the expanding community, and new rules were created to meet new needs. By 2005, three primary types of documented norms had emerged: policies, guidelines and essays. Formal norms (policies and guidelines) reflect community consensus, and can be enforced. Informal norms (essays) are not enforceable rules per se and need not reflect consensus, but do often reflect community concerns [62], and may be widely known and highly cited (such as the Bold, Revert, Discuss cycle to be discussed in section 4.3.6).

The formalization of implicit norms into rules, and the embedding of these rules in technologies such as bots and templates, facilitated distributed "peer-processes" that functioned efficiently at scale [91]. Decentralized policy creation and enforcement allowed these rules to reflect current community concerns as more editors – and, increasingly, newer editors – began to write and cite policies [6]. These findings have led researchers [91, 25] to characterize growth-era Wikipedia as an example of successful commons-based governance [71] because policies reflect local circumstances, are flexible enough to change in response to changes in needs, and are open to revision and renegotiation by the individuals who are governed by them.

No systematic analysis has been performed to track the continuation of these trends, or their impacts, into the decline period. However, evidence suggests that the rate of norm formalization has slowed. For example, decentralization has its limits: senior

editors tend to have greater “power of interpretation” over policy [48, 63] and greater control of community processes [43] than newer editors. The institution of an official peer review process for new policy proposals in 2005 may have slowed new policy creation [25]. Furthermore, recent analysis published by the Wikimedia Foundation<sup>7</sup> shows a gradual decline in participation by newer editors in the areas of Wikipedia dedicated to drafting and discussing policy, indicating that senior Wikipedians may now be more responsible for curating and interpreting community policies and guidelines than ever before.

Although these rules were originally created in order to maintain efficiency and stability in the face of a massive growth, decline-era newcomers may face entrenched social practices and processes that are no longer open to re-negotiation. If decentralization in governance and dynamic norm formalization were key to Wikipedia’s successful growth, policy calcification and centralization of policy interpretation raises concerns about the systems future viability.

**Hypothesis: Norm formalization & calcification:** *Formalization of norms has made it more difficult for newer generations of editors to shape the official rules of Wikipedia.*

As an examination of this hypothesis, I’ll report new results that demonstrate the following:

- With the introduction of a structured process for formalizing norms, the creation of new formal norms has begun to slow and the rate of rejection of contributions to formal norms has increased significantly – especially for newer editors.
- As policy creation has slowed and the rejection rate has increased, editors have begun contributing more to informal norms (essays), where their contributions are significantly less likely to be rejected.

## 4.3 METHODS

### 4.3.1 *First edit session*

To explore the reaction to newcomers during their first experience editing Wikipedia as a registered user, we borrow the concept of an edit session that was briefly discussed in chapter 2. We define an edit session as a sequence of edits performed by a registered editor to Wikipedia with less than one hour’s time between any two edits in the sequence. Given the long time some edits can take (e.g. article initiation, section writing, etc.), we expect an hour to account for time spent making an edit to an article. Note that I’ve quantitatively vetted this session cutoff other work not discussed in this dissertation [29].

---

<sup>7</sup> <http://meta.wikimedia.org/wiki/Research:WikiPride>

### 4.3.2 *Detecting rejected contributions*

Rejection of contributions in Wikipedia comes in two common forms: reverted edits and deleted edits.

A reverted edit is a contribution to an article that has been completely removed by another editor. This operation is common for removing damaging or otherwise inappropriate contributions. We use the approach described in Halfaker et al. (2009) to identify identity reverts, which restore an article to exactly the state it was in at some time before the reverted edit was made. Identity reverts are by far the most common revert type.

A deleted contribution is an edit that was made to an article that was eventually deleted. We track deleted contributions through the deleted revisions in the archive table of the MediaWiki database, so detection is trivial. In the case of newcomers, deleted edits often represent the creation of an article that is later deleted.

For both reverted and deleted edits, we limit our analysis only to encyclopedia articles since reverted and deleted contributions in other namespaces often represent different types of operations such as archiving and restructuring.

### 4.3.3 *Effect of rejection on retention*

To look for significant effects of rejection and other features of newcomer activity on retention, we apply a logistic regression over newcomers to predict a boolean metric we refer to as survival.

We define editors as surviving when they perform an edit at least two months after their first edit session. We employ an artificial sunset at 6 months such that if the surviving edit does not occur until 6 months after the first session it doesn't count. This cutoff allows us to fairly compare newcomers who started editing early in Wikipedia's history to newcomers who started up to 6 months before the end of our available data.

To examine the effects of editors' first sessions on survival, we define a set of independent variables:

- *reverted*: (Boolean) Was the editor reverted in their first session?
- *deleted*: (Boolean) Was the editor's work deleted in their first session?
- *session edits*: The number of edits completed during the first session – a proxy for an editors' initial investment in Wikipedia.
- *year*: The time at which the editor began editing in years since Wikipedia's inception (2001).
- *messaged*: (Boolean) Was the editor sent a message by another editor within the two month survival period?

- tool reverted: (Boolean) Was the editor reverted by an algorithmic tool in their first session?

#### 4.3.4 *Newcomer quality*

In order to control for the primary confounding factor in the logistic regression over editor survival – newcomer quality – we hand-coded a random sample of Wikipedia newcomers with the help of some Wikipedian volunteers<sup>8</sup>.

We randomly sampled newcomers based on when they started editing from semesters between 2001 and 2011 such that there were 100 newcomers per semester. This sampling approach allows for generating statistics for comparison over time. We built a tool for performing this qualitative analysis that allowed our coders to view a newcomer’s first session edits, but hid all information about when the edit took place to protect against a temporal bias. The tool instructed the coders to categorize newcomers into 4 ordinal categories:

1. Vandal - Editing to cause harm or offend (e.g. slurs, insults and libel).
2. Bad-faith - Damage for fun (e.g. humorous falsehoods).
3. Good-faith - Trying but not productive (e.g. non-neutral content).
4. Golden - Valuable contributions.

To check for inter-rater reliability, we produced an overlapping set by randomly sampling 100 newcomers from the primary sample to be coded by all 5 raters. The overlapping set was randomly shuffled into the work of each coder to control for an order bias. Kendall’s coefficient of concordance was lower than expected ( $W=0.413$ ,  $p<0.001$ ), so we base our results on an ordering of the two desirable categories (golden & good-faith) vs. the two undesirable categories (vandal & bad-faith). The concordance between those categories was much more respectable:

- 93.6% ratings agreed with the group
- 4.6% were too high (good rating of bad editor)
- 1.8% were too low (bad rating of good editor)

#### 4.3.5 *Tracking algorithmic tools*

In order to track the use of algorithmic tools, we employ various techniques described by Geiger et al in [32]. Due to norms around the use of such tools, we can determine whether or not algorithmic tools were used to make a contribution or reject another editor’s contribution by identifying comments left by the tool.

---

<sup>8</sup> 5 raters = 2 researchers + 3 Wikipedians

#### 4.3.6 *Conflict discussion reciprocation*

In Wikipedia, one of the most longstanding and widely cited essays is the Bold, Revert, Discuss cycle<sup>9</sup> (BRD). This essay envisions the editorial process in Wikipedia as mediated by discourse, instead of constant back-and-forth reverts (an “edit war”). Specifically, the essay states that:

1. editors ought to be bold in making whatever changes to articles they deem necessary.
2. other editors ought to be equally bold in reverting those changes if they do not approve.
3. upon being reverted, the original editor should use the article’s talk page to discuss the change with others, most notably the editor who reverted the change.

Both Wikipedians and researchers of Wikipedia have argued that article talk pages are a critical aspect of how content is negotiated in Wikipedia[91, 78]. To explore our intuition that editors using algorithmic tools would reciprocate at lower rates than those who were not using tools, we performed the following analysis of the BRD cycle. First, we identified every instance of the first three elements constituting the BRD cycle: an editor making a change to an article, another editor reverting that change within 14 days, and the first editor writing to the article’s talk page in response. If the reverted editor made a post to the article’s talk page within 7 days, we classified that as an initiation. We then examined future comments in the article’s talk page to see if the editor who made the revert responded to the talk page post within 7 days. If the reverting editor made a post to the talk page, we classified that as a reciprocation.

Because this analysis was done algorithmically, reciprocation may be over-represented if, for example, the reverting editor responded to a different post and ignored the post by the reverted editor. Since we hypothesize lower rates of reciprocation, this possible over-representation was deemed acceptable. To minimize cases of talk page vandalism or counter-vandalism appearing like a BRD initiation/reciprocation, we disregarded any talk page posts that were either reverted within 12 hours or that were themselves reverts of earlier revisions. Because we were interested in how tools are affecting the relationship between new and veteran editors, we only looked at cases in which the reverting editor had been registered for over 30 days and the reverted editor had been registered for under 30 days.

#### 4.3.7 *Policy growth and calcification*

In order to examine the activity surrounding norm formalization in Wikipedia, we used the category hierarchy to identify the pages considered to be policies, guidelines

---

<sup>9</sup> <http://en.wikipedia.org/wiki/WP:BRD>

and essays. To measure the growth of norms over time, we use a set of metrics to track activity in norm pages.

- contributors: The number of registered editors that contributed to norm pages.
- contributions: The number of contributions to pages in a norm category.
- length change: The change to the overall length of pages in a norm category.

To look for evidence of calcification we used a logistic regression over the Boolean outcome of whether a contribution to a norm page was reverted. We define a set of independent variables:

- editor tenure: The age of an editor in years since account registration
- year: The time in years since Wikipedia’s inception (2001)
- essay: (Boolean) Is the page an essay?

To identify policy proposals, we performed a text analysis on a diff dataset<sup>10</sup> published by the Wikimedia Foundation. Using the dataset, we tracked additions and removals of the “proposed” template to determine when pages were nominated for the formalization process.

I assumed that pages currently categorized as policies or guidelines were formalized while pages outside of those categories were not.

## 4.4 REJECTION & RETENTION

### 4.4.1 Results

To explore the validity of **HYP Rejection & retention**, we first looked for a significant relationship between rejected edits and survival. As described in Methods, we use a logistic regression over the first session edits to determine the likely effects of various first edit session metrics.

The “All” newcomers column of table 6 shows a significant negative effect for both editors who were reverted or had their revisions deleted in the first edit session. This result supports our hypothesis and re-affirms the conclusion of chapter 3 that reverts of newcomer’s contributions reduces the rate of survival. The regression also reports a significant negative effect for year. This suggests that while rejection is a strong negative predictor for survival, there are other independent effects over time that are reducing the rate of survival for newcomers.

However, these results alone do not represent a good test of **HYP Rejection & retention** since vandals and other unwanted editors could represent the rejected and non-surviving editors. To explore this confound, we turn to our analysis of the quality of newcomers.

<sup>10</sup> <http://dumps.wikimedia.org/other/diffdb>

Table 6: The coefficients of a logistic regression over the first edit session of two sets of randomly sampled Wikipedia users predicting survival are presented. All newcomers represents a purely random sample of registered users from Wikipedia. Desirable newcomers represents the subset of editors sampled for quality analysis that were determined to be at least acting in good-faith.

	All (n=100k)			Desirable (n=1708)		
	$\beta$	SE	$P(>  z )$	$\beta$	SE	$P(>  t )$
(Intercept)	-1.98	0.017	< .001	-1.30	0.098	< .001
year	-0.40	0.012	< .001	-0.59	0.069	< .001
session edits	0.18	0.009	< .001	0.24	0.064	< .001
deleted	-1.45	0.037	< .001	-0.80	0.217	< .001
reverted	-0.68	0.035	< .001	-0.50	0.173	.004
messaged	0.54	0.027	< .001	0.68	0.127	< .001
tool revert	-0.67	0.062	< .001	-2.16	1.086	.047

Figure 18 shows that, while the combined proportion of newcomers falling into the two desirable categories fell from 92.2% in the first semester of 2005 to 79.8% in the first semester of 2006, the combined proportion of desirable newcomers stays relatively consistent from 2006 forward. Notably, this shift to a new consistency in 2006 occurred about one year prior to the peak and decline in Wikipedia's active contributors that began in 2007 (see figure 17).

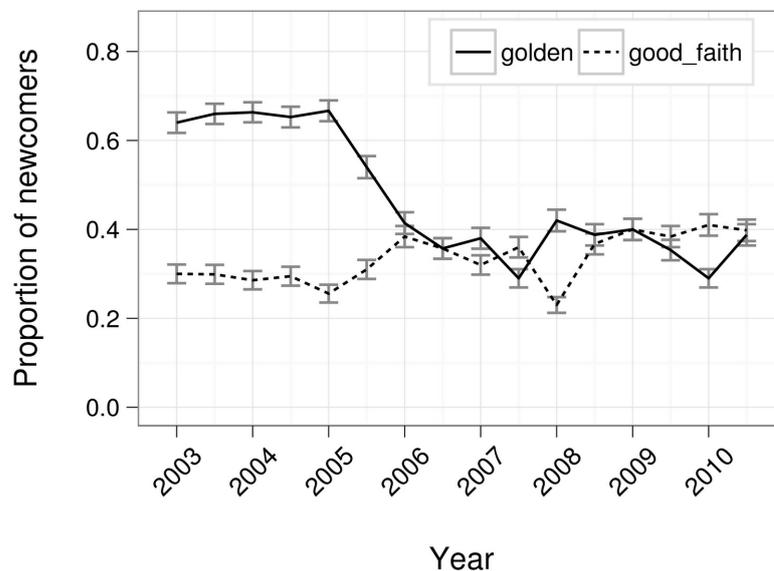


Figure 18: **Quality of newcomers over time.** The proportion of editors falling into the two good-faith quality categories is plotted over time.

Figure 19 shows a general increase in the rate of rejection for desirable newcomers over time. As hypothesized, the rate of rejection rises substantially for good-faith editors (editors who appear to be trying to be productive, but unsuccessful). The most

substantial change to the rate of rejection of desirable newcomers occurred during the time between the first semester of 2006 and the first semester of 2007 (during transition from growth to decline). We observed a shift of 6.1% to 18.2% of desirable newcomers experiencing rejection in the form of a revert.

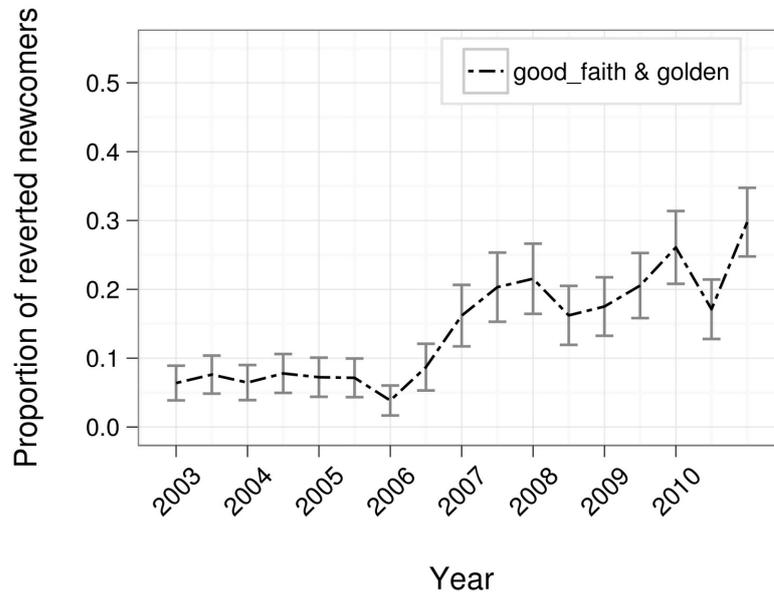


Figure 19: **Reverts of desirable newcomer contributions over time.** The proportion of good (“good-faith” & “golden” combined) newcomers with at least one reverted first session edit is plotted over time.

As figure 20 shows, a decline in the survival rate for desirable newcomers corresponds closely with the increasing revert rate observed in figure 19. Again we see the most substantial shift occurring during the timespan that Wikipedia’s editing community transitioned from growth to decline. In the first semester of 2006, 25.6% of desirable newcomers continued editing for at least two months. Within a year, the desirable newcomer survival rate falls to 11.7% and does not recover.

To determine if the rejection of first session contributions has the same effect on desirable newcomers as it does on overall newcomers, we performed a similar regression to predict survival over only the desirable newcomers. Table 1 shows that each one of the predictors affects all newcomers and desirable newcomers in the same direction.

These results support our hypothesis. It appears that the rising rate of rejection of newcomers’ first session contributions is predictive of the decrease in newcomer retention.

#### 4.4.2 Discussion

Our results suggest that rejection of contributions, especially for desirable newcomers, has substantially affected the decline. In both of our regressions, rejection in the forms of both reverted and deleted contributions to articles were independently significant

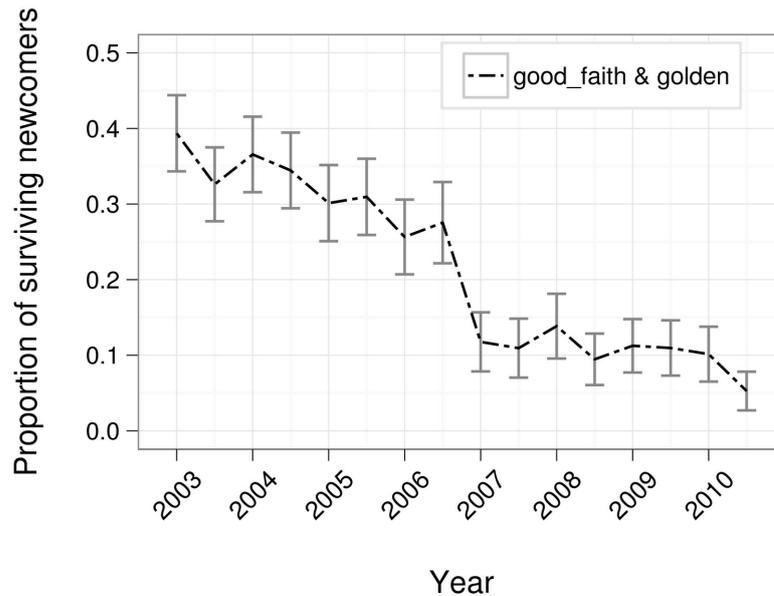


Figure 20: **Survival of desirable newcomers over time.** The proportion of surviving good (“good-faith” & “golden” combined) newcomers is plotted over time.

predictors of the retention of desirable newcomers. Rejection is reported to be a significant predictor of retention independent of the age of the project. This means that rejection was likely to be a demotivator to newcomers who joined the project long before retention of newcomers became an issue.

We also found that over the lifetime of Wikipedia the probability that contributions made by desirable newcomers are rejected has increased. Our impression from the qualitative hand coding of newcomer quality is that, the majority of the time, these rejections were due to misunderstandings about the norms of the community. This result suggests that “unwanted” but not intentionally damaging contributions may have been handled differently in the past.

One such way of dealing with imperfect contributions without sacrificing quality is to “massage” them into a form that is valuable for an article. Perhaps the increasing use of tools that afford only two possible reactions – *accept* or *reject* – are making it more likely that contributions are rejected outright.

## 4.5 TOOL USE AND CONSEQUENCES

### 4.5.1 Results

**Newcomer rejection.** To explore the potential role of algorithmic tools as gatekeepers to the community, I examine the interactions between newcomers and the actions performed via algorithmic tools. Figure 21 shows the growing use of algorithmic tools to reject the contributions of newcomers in Wikipedia. The plot shows that, around the beginning of exponential growth, which is the same time that the first algorithmic

tools for rejecting contributions were released, the proportion of newcomer contributions that were rejected using tools rose to 30%.

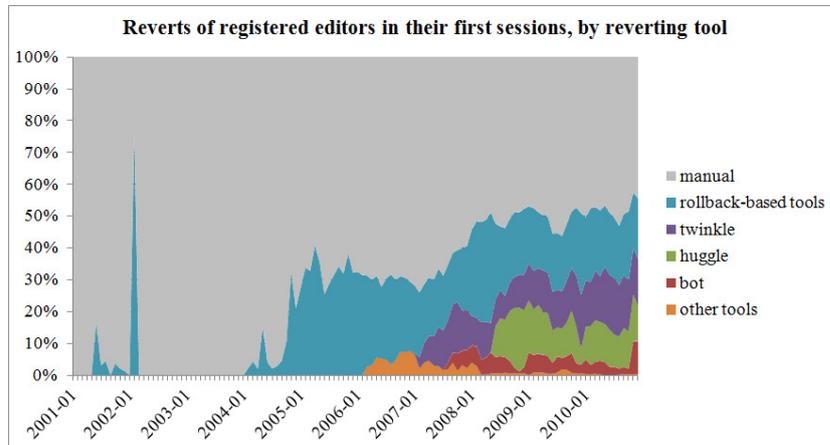


Figure 21: **Use of algorithmic tools to reject newcomers edits.** The proportion of rejected first session contributions is plotted for newcomers by the mechanism used for rejection over time.

The majority of tool-based rejection of newcomers came from human-computation tools – tools that borrowed human judgment. This seems reasonable given that there were several early controversies regarding the way registered editors were treated by bots that resulted in a normative framework that forced bot developers to tread lightly when dealing with community members [27].

**Discussion reciprocation.** For editors who revert manually, the rate of reciprocation has dropped slightly, from a peak of 67% in 2005 to 56% in 2010. The overall rate of reciprocation has dropped dramatically, since none of the major bots are programmed to reciprocate BRD initiations.

Most striking is the rate of reciprocation by users of Huggle, a standalone program that is designed specifically to allow humans to judge and revert edits as fast as possible. Editors who revert using Huggle have an average response rate of 7%, compared to editors who use the browser-based extension Twinkle, which has an average response rate of 53% – only slightly lower than editors who revert manually.

The rollback feature is a sort of confluence of different revert tools since it can be used in the browser as well as in a variety of plugins and standalone programs to revert content en masse. Users of rollback show a rate of reciprocation around 30% – this is in between Huggle and Twinkle, likely due to the many different ways in which the functionality is accessed.

**Rejection & retention.** To explore whether rejection via algorithmic tools is a significant predictor for survival in Wikipedia, we included a Boolean independent variable in the regressions described in table 6. Both columns report a significant negative effect for tool based reverts on the survival of newcomers. This result suggests that reverts of desirable newcomer contributions by Wikipedians using automated tools exacerbate the negative effect of rejection on survival.

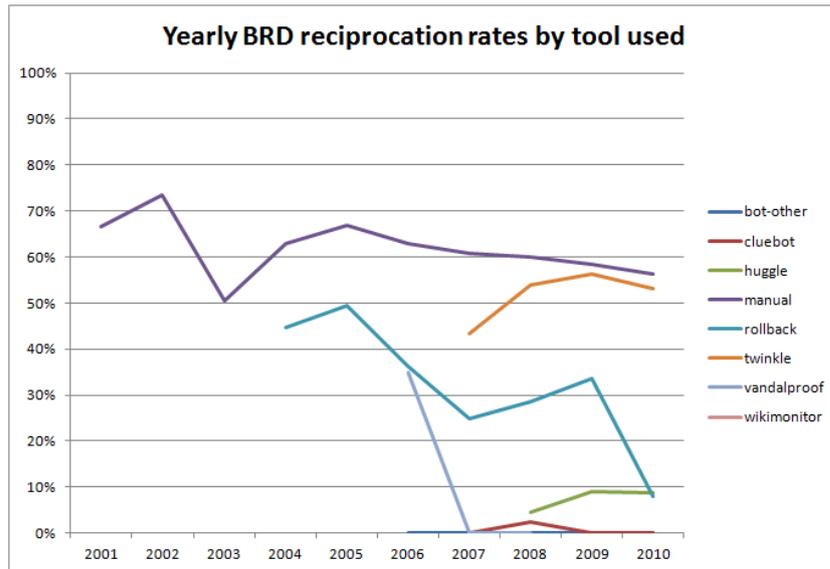


Figure 22: **Use of algorithmic tools to reject newcomers edits.** The proportion of rejected first session contributions is plotted for newcomers by the mechanism used for rejection over time.

Since the exponential growth of Wikipedia, the rate at which desirable newcomers are reverted using tools also appears to be rising. Figure 23 shows a rise in the rate of tool based rejection of newcomer contributions from 0% in 2006 to 40% in 2010.

#### 4.5.2 Discussion

Our analysis shows that algorithmic tools have had an increasing role in rejecting the contributions of newcomers. Given that my work with Geiger showed that these tools are also taking over the task of “welcoming” newcomers via warning messages posted on their talk page [32], this suggests that newcomers are increasingly rejected by and warned by these not-entirely-human actors. Our results also show that when these newcomers attempt to interact with Huggle users through the community’s preferred approach about their rejected contributions, they tend to be ignored. It appears that Wikipedia’s gatekeeping practice is shifting from human, personal interaction to mechanical, impersonal interaction and the switch primarily took place at the end of the exponential growth period for the community.

The regression analysis over survival (presented in table 6) shows a significant, exacerbating effect for the newcomers whose contributions were rejected using tools. The discussion reciprocation analysis showcases one instance in which tool users are generally not interacting in a way that we assume would be positive and helpful to newcomers. Overall, I suspect that this impersonal, noncommunicative nature of interaction has other, possibly more difficult to measure, implications that are exacerbating the effect of rejection on retention.

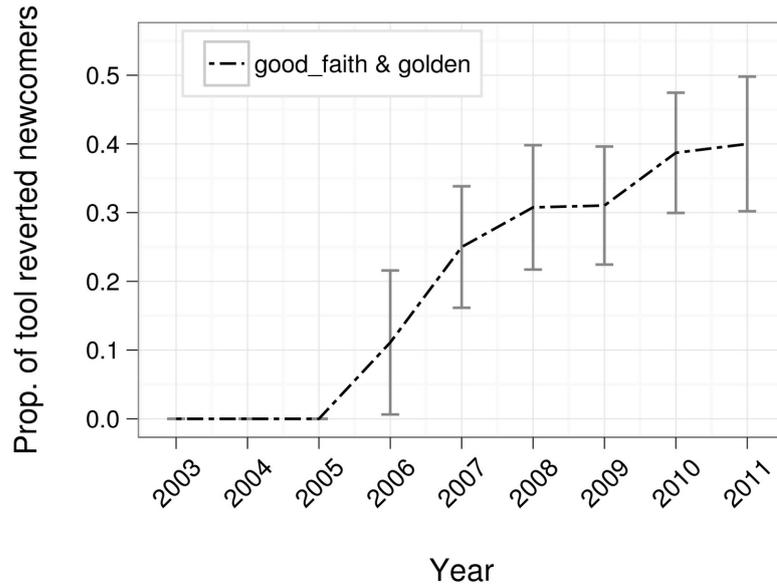


Figure 23: **Rate of automated reverts for desirable newcomers.** The proportion of reverted desirable newcomers (“good-faith” & “golden” combined) who were reverted using algorithmic tools is plotted over time.

## 4.6 NORM FORMALIZATION AND CALCIFICATION

### 4.6.1 Results

To explore **HYP Norm formalization & calcification**, we first looked for changes in the rate of new policy creation following the introduction of a structured proposal process in 2005. Figure 24 shows that growth of policies and guidelines began to slow in 2006, just as Forte et al. reports [25]. The results from our analysis of new policy/guideline proposals show that the number of new policy proposals accepted via this process peaked in 2005 at 27 out of 217 (12% acceptance). 2006 saw an even higher number of proposed policies, but lower acceptance with 24 out of 348 proposals accepted (7% acceptance). From 2007 forward, the rate at which policies are proposed decreases monotonically down to a mere 16 in 2011 while the acceptance rate stays steady at about 7.5%.

Existing formal norms continued to be revised and expanded through 2006, which closely correlates with the end of the community growth period (see figure 17). After that point, growth existing policies and guidelines begins to decline.

To look for the effects of policy calcification, I compared the rate of creation and contribution to formal norms (policies and guidelines) and informal norms (essays). A rise in the rate of essay creation corresponds to the decline in policy and guideline creation. 69 essays were written in 2005, 164 in 2006 and the rate doesn’t fall below 185/year thereafter. This initial growth in new essays appears to be due in part to the conversion of failed policy/guideline proposals: in 2006, 22% of new essays began as

failed policy proposals.

However, the percentage of essays that started out as rejected policies or guidelines decreases sharply to 12% in 2007 and 1% by 2011. Figure 24 shows that the growth of essays overtakes both policies and guidelines in 2006 and continues to rise to 1.52 MB of new content per year by 2008. From that point forward, the volume of content contributed to essays remains consistently above policies and guidelines. The number of distinct contributors to essays over time (not shown) follows a similar pattern.

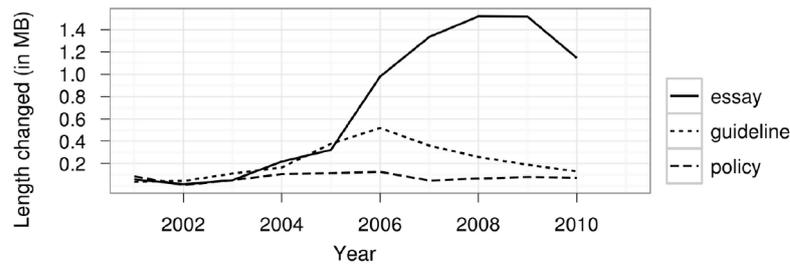


Figure 24: **Norm page growth over time.** The change to overall size of the three norm types is plotted by year.

To look for evidence of calcification of policies against contributions, we performed a logistic regression (described in section 4.3) to predict the rejection of new contributions to all three types of formalized norms. Table 7 shows a significant, positive effect for the year in which contributions were made which suggests that, over time, contributions to all types are more likely to be rejected independent of the tenure of the editor making the contribution.

However, the regression also reports a significant negative interaction between the year in which the contribution was made and the Boolean variable that codes for essays with a coefficient at a comparable scale (-0.12 vs. 0.10). This suggests that, for essays, the increasing rate of rejection is almost entirely negated. The significant, negative effect reported for the editor's age (tenure) suggests that more senior editors are less likely to have their contributions to norms rejected in general, but again we see a reversed effect with the interaction with essay (-0.29 vs. 0.06). This suggests that newer editors are significantly more likely to be successful when contributing to essays.

#### 4.6.2 Discussion

Our analysis shows that the growth of new formal norms has declined, and it has become more difficult over time for Wikipedia editors to change the existing rules – especially for new editors. The rising rate of rejection is evidence of calcification and the slowing growth of formal norms is an inevitable outcome of such a process.

Two consequences of calcification bear directly on newcomer socialization and retention. First, the effects of calcification are disproportionately felt by newer editors who

Table 7: The coefficients of a logistic regression over the contributions of registered editors to norm pages predicting success (i.e. not reverted) are presented.

(n=120535, AIC=16801)			
	$\beta$	SE	P(>  z )
(Intercept)	-1.50	0.043	< .001
editor tenure	-0.29	0.006	< .001
year	0.12	0.006	< .001
essay	-0.38	0.135	.005
editor tenure:essay	0.06	0.019	.002
year:essay	-0.10	0.019	< .001

see their edits to policies and guidelines rejected at a higher rate. This suggests that under Wikipedia’s current regime, rules are less open to revision by affected editors than they were during the growth period, decreasing the dynamic flexibility that was key to Wikipedia’s adaptive success. Second, although newer editors are contributing more to essays – where their contributions are less likely to be reverted – essays are not official, enforceable rules and are not widely cited. While an increase in essay writing is an encouraging sign of newer editors’ continued interest in participating in community governance, it is not an effective mechanism for social change.

#### 4.7 CONCLUSION

Wikipedia has changed from “the encyclopedia that anyone can edit” to “the encyclopedia that anyone who understands the norms, socializes him or herself, dodges the impersonal wall of semi-automated rejection and still wants to voluntarily contribute his or her time and energy can edit”.

Rejection of unwanted contributions is Wikipedia’s primary quality control mechanism [82] and it works [33]. However, as the scale of the system has increased, rejection of newcomer contributions has increased, with the unintended consequence of driving away well-meaning newcomers. However, outright rejection of a contribution isn’t the only way to control quality. A contribution that adds some type of value, but possibly in the wrong context, location or format can be accepted via a rewrite. We suspect that the growing use of algorithmic tools may have affected a transition from rewrites to reverts due to the fact that these tools often only afford the decision of “accept” or “reject”.

However, these tools were instrumental in improving the efficiency and effectiveness of managing damage and deviant users [30, 28]. Without algorithmic tools, substantially more volunteer effort would be needed to protect the encyclopedia from damage, and system efficiency would likely suffer.

Even newcomers who make it through their initial contributions are encountering resistance while attempting to enter Wikipedia’s inner circle. While Wikipedia suc-

cessfully democratized policy creation and enforcement during the time of exponential growth, we've shown that the community's artifacts of governance have calcified, making rules less adaptable and harder to change, especially for new editors. These editors increasingly appear to be moving to less formal spaces to construct and discuss ideas about Wikipedia's goals, processes and organization. However, lacking the exposure and enforceability of policy, these contributions are unlikely to gain wide currency within the community, shift community norms around interacting with newcomers, or help the community tackle issues related to the editor decline.

Wikipedia's challenges may seem unique to its status as one of the largest collaborative projects in human history, but the widespread use of algorithmic tools to maintain social order online makes Wikipedia's response quite relevant to a variety of other production systems. For example, Lampe and Resnick studied the distributed system of meta-moderation and "karma" used in Slashdot to remove inappropriate comments and bring the most interesting and insightful commenters to the top of a discussion thread [53]. Another study by Gillespie examined the copyright infringement detection algorithms used by YouTube to automate the process of identifying and removing infringing content[86].

While concerns surrounding new user retention are not as immediately pressing for those two websites as for Wikipedia, they show two alternative responses to the various issues that arise in mediating participation online. In general, the case of Wikipedia shows how, in all mediated platforms, designers, managers, and community members must think about the relationship between the tools that social systems use for enforcement and the kinds of activities that those tools afford and restrict.

#### 4.7.1 *Recommendations*

During the growth of contributors in Wikipedia's community, volunteers built effective automated tools for (1) specifying what does and doesn't belong in the encyclopedia and (2) rejecting those contributions (and contributors) that don't belong.

Today, Wikipedians need good tools to help them identify and support desirable newcomers who get caught in the crossfire. Luckily, the most difficult part of this problem has already been solved. Existing intelligent algorithms effectively detect the edits of undesirable newcomers, so the converse problem is already well defined and has a set of candidate solutions.

Once desirable newcomers are detected, Wikipedia should develop effective mechanisms to reach out to them. The ways of reaching out that are most reliably successful are human-to-human: an experienced Wikipedia contacting a newcomer to establish a supportive relationship. However, the current mentoring system in Wikipedia has been relatively unsuccessful, and unable to achieve scale [66]. Wikipedia should explore novel ways of reaching out effectively to its most valuable resource: desirable newcomers.

## 5 EXPANDING PARTICIPATION AT THE PERIPHERY

---

### 5.1 INTRODUCTION

As I argued in chapter 4, open production systems like Wikipedia require a stable pool of volunteer contributors to remain productive. Without volunteers to occupy necessary roles, these systems would cease to function. The success of an open production system appears to be highly correlated with the number of participants it maintains[? 18]. In order to maintain the pool of contributors, newcomers must be continually socialized into the organization[85, 73].

#### 5.1.1 *Participation in peer production*

Participation in online communities tends to manifest as a long-tail distribution[100]; a tiny, active minority produces most of the content while the majority of community members produces very little individually. Beyond the contributing population is an often overlooked population of users who do not contribute, commonly referred to as “lurkers”. Despite the fact that lurkers do not contribute, their numbers tend to dwarf the rest of the community by orders of magnitude. Recent research suggests that the primary reason for lurking in online communities is a lack of a perceived “information benefit” in increasing their effort to contribute[69, 10, 42], yet in some communities, a substantial proportion of lurkers are simply not aware that they can participate[3, 68].

Recent work has shown that the English Wikipedia, an online encyclopedia often held as a prototypical example of open collaboration, has a particularly steep long-tailed distribution of participation[100]. Through an analysis of Wikipedia’s historical edit logs, Priedhorsky et al. estimated that the most active 0.1% of contributors produce nearly half of the encyclopedia’s value[74]. According to the Wikimedia Foundation’s official statistics<sup>1</sup> report for March of 2012, the encyclopedia was edited by 113,304 editors and a comScore<sup>2</sup> report for the same month shows 1.47 billion unique visitors. These numbers suggest that the English Wikipedia’s consumers outnumber producers by 10,000 to 1. Given what the current contributors have been able to achieve – a vast and highly accurate encyclopedia – the consumers of Wikipedia would represent a massive potential workforce if even a small percentage of them could be coerced to contribute productively.

Research in collaborative computing has explored effective mechanisms for boosting participation in online communities. For example, Beenen et al. used insights from the collective effort model (CEM) to increase participation in a movie recom-

---

<sup>1</sup> <http://stats.wikimedia.org/EN/TablesWikipediaEN.htm>

<sup>2</sup> <http://comscore.com>

mender via email requests[5]. Rashid et al. extended their study by replacing the email requests with visual queues within the user interface[75]. Within Wikipedia, Cosley et al. showed the effectiveness of a task routing system designed to decrease the cost of finding *where* to contribute in order to boost contributions of current Wikipedians[15]. Wash and Lampe showed that directly asking users for comments on news articles could temporarily increase participation without reducing the quality of comments[95]. However, none of these studies examined the process of transitioning from non-participant to participant.

### 5.1.2 *Supporting peripheral participation*

Social learning theory gives us a framework for understanding how newcomers to a community transition from non-participation to participation – i.e. from consumer to producer. In their highly cited work on how newcomers approach “communities of practice”, Lave & Wenger describe the process by which initiates begin participation on the periphery of a community by performing simple and low-risk yet productive tasks that they refer to as legitimate peripheral participation (LPP)[55]. As newcomers gain experience, they become more familiar with the tasks, vocabulary and norms of the community, and by doing so, they are able to confidently expand their level of participation. Recent work by Preece & Shneiderman applies this framework to online communities by defining a set of roles that users adopt as they transition from the periphery of the community (reader) to its center (leader). Specifically, they recommend that online communities should be designed to “support legitimate peripheral participation so that readers can gradually edge into contributing”[73].

The open source software community has engineered a model for supporting a type peripheral participation in the form of open issue tracking software. Through such systems, users who do not have the knowledge, time or interest in directly contributing code to a project may submit bug reports and feature requests to the developers. Through the use of such systems, participation is expanded from a relatively small number of producers (developers) to a much more numerous group of consumers. This contribution medium allows these consumers to learn about the norms and practices of a community by interacting at the periphery. In an analysis of the growth of the developer community around the Freenet project, an open source file sharing platform, Von Krogh et al. describes a “joining script” where some newcomers post bug reports and feature requests before attempting substantial participation. Newcomers who follow this sequence of activities are more likely to be granted privileged access to the source code repository[92].

Such expanded participation models allow prospective contributors to make simple and low-risk yet productive contributions, and by doing so, the workload for maintaining the community product is distributed more evenly among its stakeholders. When such participation is high in quality, it can reduce the workload of the small percentage of prolific contributors by spreading a time-consuming production activity

(e.g. bug detection and reporting) over more individuals. This reduction in workload for the primary contributors is particularly valuable for volunteer communities where effort is the primary currency of progress.

When viewed this way, Wikipedia is not keeping pace with developments in collaborative computing that leverage peripheral participation. Although Bryant et al. uncovered, through a series of interviews with highly active Wikipedia editors, that social learning theory is consistent with the way that editors view their integration into the Wikipedia community[9], the most peripheral type of activity they describe is making edits to fix mistakes. Antin and Cheshire argue that reading the encyclopedia should be viewed as a form a legitimate participation by showing evidence that experienced readers of Wikipedia know more details about the editing community[3], but they do not explain how reading is “productive”, a condition of legitimate peripheral participation as described by Lave & Wenger[55]. MediaWiki, the software that runs Wikipedia, affords no half-step from reader to editor and my work suggests newcomers are finding it increasingly difficult to transition into productive editors. Given that lurkers in other communities have reported fear of strong, negative reactions as their reason for not contributing [42], it seems likely that many Wikipedia readers don’t contribute due to a reasonable fear of rejection.

However, opening up peripheral participation to a larger pool of users doesn’t come for free. The net value of expanding participation often depends on the associated cost of filtering and moderating a larger volume of contributions. A recent analyses of bug reports by casual users in large-scale open source software projects suggest that the mismatch between what users report and what developers find useful can undermine the value of broader participation [7]. A related study reports that core developers in software projects find higher value in smaller groups of highly committed bug reporters than in a larger group of unengaged contributors [46]. Given these concerns, it is essential to consider both a potential increase in participation and the moderation concerns when vetting methods for extending participation.

### 5.1.3 *Article feedback*

Motivated by decreasing levels of newcomer retention in the English Wikipedia’s contributor community[99, 84], I worked with the Wikimedia Foundation to develop a half-step between reading and editing in the form of an extension to the Wikipedia’s software called the *Article Feedback Tool* (AFT). This tool allows readers to submit feedback about encyclopedia articles to editors via the standard web interface. Like bug tracking software, AFT allows the community’s consumers to communicate their concerns to the community’s producers in a simple, low risk way. Submitting feedback to an article’s editors can also be productive; like bug reports, feedback can be used by the producers as a mechanism for identifying problems and missed opportunities in articles. In this way AFT can be viewed as an extension of the contribution model of Wikipedia that supports a new mode of legitimate peripheral participation.

In this chapter I'll extend the online participation literature and examine an aspect of LPP in Wikipedia through a set of field experiments (performed live on the English Wikipedia) designed to test AFT's effectiveness in eliciting participation from readers and encouraging the conversion from reader to editor. I'll also analyze the utility and productivity of the increased participation in the context of Wikipedia's quality control mechanisms.

## 5.2 THE ARTICLE FEEDBACK TOOL

The *Article Feedback Tool*<sup>3</sup> is an extension of Wikipedia's user interface that affords readers of an encyclopedia article the ability to submit feedback to editors. When a reader views an article, a small form appears after the content asking the reader to "Help improve this article". Figure 25 displays the different versions of this form that were tested as part of this study. Each form consists of two components: a question prompt, which I refer to as an "elicitation", followed by a free-form text box. The feedback forms differ only by their elicitations:

- Form 1 asks the reader if she found what she was looking for: yes or no.
- Form 2 allows the reader to pick from 4 types of feedback she might give in the text box: suggestion, praise, problem or question.
- Form 3 asks the reader to rate the article on a 1-5 scale.

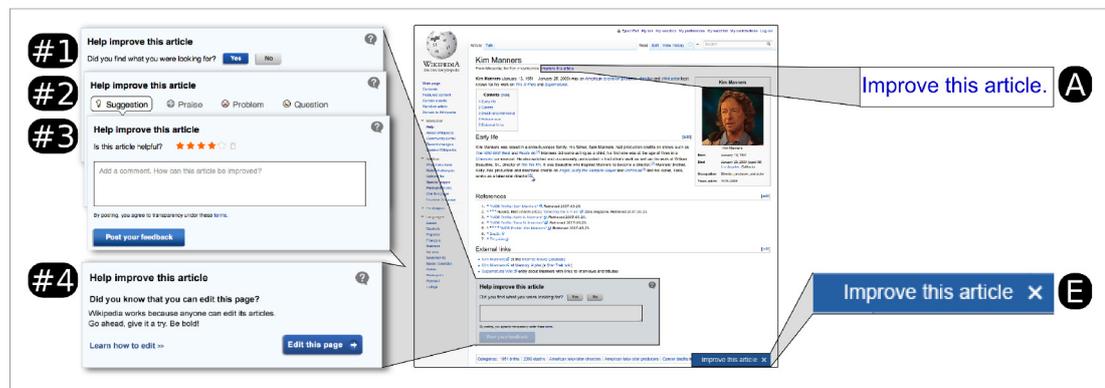


Figure 25: **The Article Feedback Tool's interface components.** The components of the AFT interface are called out from Wikipedia's article viewing interface. An article on Kim Manners, one of the randomly sampled articles, is loaded. #1-3 represent different versions of the article feedback forms. #4 represents the edit invitation form, a request for the reader to try editing the page. A and E represent links inserted into the page to direct readers to the feedback form.

To evaluate the feedback form's effect on the conversion from reader to editor, I introduced a fourth type of form for comparison. Form 4 doesn't accept feedback; instead, the form invites the reader to make a contribution by editing the article using a UI of exactly the same size and location as the feedback forms.

<sup>3</sup> [http://www.mediawiki.org/wiki/Article\\_feedback](http://www.mediawiki.org/wiki/Article_feedback)

Since many Wikipedia articles are very long, the feedback forms are often hidden beneath pages of scrolling. To allow for more straightforward access to the form and to encourage more readers to leave feedback, I tested two prominent links (labeled A and E in figure 25) that act as shortcuts to the form. When a reader clicks on a prominent link, the form is loaded as an overlay on the page – allowing a reader to access the form without having to scroll. Although Link A appears at the top of the article, it is relatively hidden by the boilerplate statement “From Wikipedia, the free encyclopedia”. Link E is intended to be much more likely to catch a reader’s attention by (1) taking up more space, (2) making clear that it is a button and (3) remaining visible to the reader by staying fixed to the bottom of the window as she scrolls through the article.

When performing the experiments (described in section 5.5), I mixed and matched these interface components to form experimental conditions that test the effects that each component has on the quality and quantity of participation.

### 5.3 RESEARCH QUESTIONS

In this section, I’ll motivate three research questions that drive my analysis and exploration of the effectiveness of AFT as a mechanism for eliciting useful feedback and converting readers to Wikipedia editors.

#### 5.3.1 *RQ1: How do different elicitations affect the volume and utility of feedback?*

A larger pool of contributors isn’t necessarily better. Although Wash and Lampe found that increasing participation by asking readers for comments on news articles did not decrease comment quality[95], these comments were not judged for their *usefulness* to the journalists. When Bettenburg et al. measured the usefulness of bug reports to developers, they found that reports by casual users were less useful due to mismatches between user and developer concerns[7]. We hypothesize that focusing readers toward the concerns of editors when eliciting feedback will encourage submissions that are more useful to editors. Conversely, we expect designing the feedback form around the concerns of readers may encourage more readers to submit feedback at the cost of decreased utility to editors.

We expect to find a tradeoff between the quantity and quality of participation whereby elicitation that receive more feedback submissions will do so by encouraging submissions that are less useful to editors, or more precisely:

- Asking readers whether they found what they were looking for (form 1) should increase participation since the question directly addresses readers’ concerns but decrease utility since many of those concerns will not be shared by editors.
- Asking readers to categorize their feedback (form 2) should elicit more useful feedback by giving cues to readers about what types of feedback are expected

but decrease participation due to the increased transaction cost of requiring the reader to provide meta-information.

- Asking readers to rate the article (form 3) should increase the quality of comments by encouraging readers to consider the the article’s quality within the context of the encyclopedia (editors’ concern) but decrease participation by forcing readers into the role of experts.

**Hypothesis: Participation & utility:** *Form elicitation that increase participation will do so at the cost of the decreased utility of submissions.*

### 5.3.2 RQ2: *How does the prominence of the elicitation affect the volume and utility of feedback?*

The design decision to place the feedback forms after an article was strategic. Given that most articles in Wikipedia contain enough content to require scrolling the browser window, feedback forms positioned at the bottom of the article are hidden such that only readers who have scrolled through the article will ever see them appear on the screen. Presumably, these readers are particularly suited to leaving feedback because they should be more likely to have examined some content from the article before seeing the form. However, there may be readers who would have submitted valuable feedback, but never realized they could because they didn’t scroll through the entire article.

By presenting a prominent link to the feedback form at the top of the article, we provide an alternate route to leaving feedback that does not require the reader to scroll through the article and we expect that the prominent placement will capture the attention of more readers. We expect that by making feedback forms more prominent in this way, we will increase the number of submissions at the cost of decreasing the overall usefulness of submissions.

**Hypothesis: Prominence & volume:** *The volume of feedback submissions will increase with prominence.*

**Hypothesis: Prominence & utility:** *The utility of feedback submissions will decrease with prominence.*

### 5.3.3 RQ3: *How does the presence of the feedback interface affect new editor conversion?*

A prominent invitation to leave feedback as opposed to making a productive edit to an article may undermine the “sofixit”<sup>4</sup> and “be bold”<sup>5</sup> culture of Wikipedia that encourages individuals who feel that there’s something wrong with an article to boldly make the edit themselves. Further, even for a reader whose interest in making an edit

<sup>4</sup> <http://en.wikipedia.org/wiki/Template:sofixit>

<sup>5</sup> <http://en.wikipedia.org/wiki/WP:BOLD>

herself was held constant despite submitting feedback, the act of submitting feedback should “cannibalize” some of her finite time and effort that could have been put towards more direct participation in the form of an edit. If this is true, the presence of AFT may actually be counterproductive.

**Hypothesis: Conversion:** *The presence of AFT will decrease the rate of new editor conversions.*

As a stopgap measure, presenting the reader with an invitation to edit (form 4 in figure 25) after submitting feedback should encourage potential new editors to continue through the normal process of “boldly” making an edit themselves. However, such an invitation to edit might also short circuit the normal process of peripheral participation by encouraging a reader to make the transition to editing before they’ve lurked for long enough to understand the norms and goals of the community.

**Hypothesis: Converted productivity:** *New editors who are invited to start editing articles will be less productive than editors who start by their own volition.*

## 5.4 METHODS

### 5.4.1 *The article sample*

To explore our three research questions, we performed three distinct experiments on the English Wikipedia. For each experiment, the AFT interface components were loaded on a 0.6% random sample of encyclopedia articles. All readers who viewed these articles were randomly assigned to experimental conditions (described in section 5.5). We used browser cookies to extend the continuity of experimental groups between sessions and internet connections (IP addresses) so that readers would remain within the same experimental condition through the duration of an experiment.

To minimize the potentially disruptive impact of testing AFT on a live website, the same random sample of articles was used for all three experiments. However all users were re-bucketed between experiments to control for a potential ordering bias.

### 5.4.2 *Feedback utility*

In order to address the research questions postulated in section 5.3, we required a way to determine which of the feedback submissions were useful to editors. The judgement of experienced Wikipedia editors to perform this evaluation was invaluable in determining the usefulness of feedback since, by design, they will be responsible for making use of feedback after it has been submitted, so they will know best which feedback submissions are useful. In other words, we consider their judgement to be a direct measure of utility. I organized a group of Wikipedian volunteers interested in determining the overall utility of AFT via requests posted on the documentation

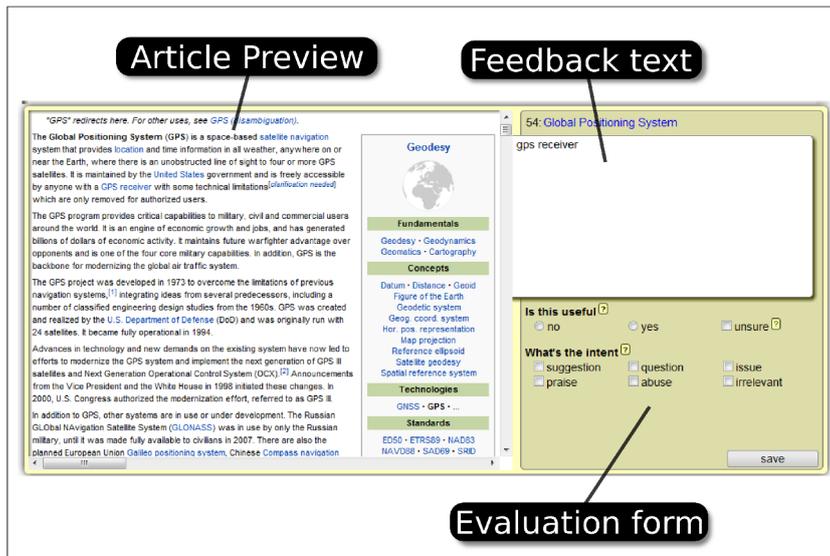


Figure 26: **The Feedback Evaluation System.** FES, the interface used by ikepedians to evaluate the usefulness and type of feedback submitted is resented with annotations for the three main components.

page<sup>6</sup>.

To support these Wikipedian evaluators, I built a user interface (see figure 26) to present a feedback submission in the context in which it was submitted while hiding details about the experimental condition. For each feedback submission loaded into this user interface, the version of the article at the time feedback was submitted is loaded into an article preview pane, and the feedback text that was submitted is loaded into the feedback text pane. The rater was instructed to use an evaluation form to answer two questions:

**Is this useful?.** Raters were instructed to mark a feedback submission as “useful” if they could imagine making use of this feedback to edit the article<sup>7</sup>.

**What’s the intent?.** Raters were instructed to categorize the intent of the feedback. They could select zero to many options from: *suggestion*, *praise*, *question*, *issue*, *irrelevant* and *abuse*. These categories were decided upon through a pilot run in cooperation between the researchers and a subset of the raters.

20 Wikipedians participated in this hand-coding process. We randomly assigned work such that each feedback submission was rated by exactly two different Wikipedians. Each Wikipedian rated between 50 and 350 feedback submissions.

### 5.4.3 New editor productivity

To measure the productivity of newly converted editors, we examine their contributions to encyclopedia articles during their first week of tenure. We assume that an edit is “productive” if it is not reverted by another editor within 48 hours.

<sup>6</sup> [http://en.wikipedia.org/wiki/Wikipedia:Article\\_Feedback\\_Tool](http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool)

<sup>7</sup> Raters were allowed to select a checkbox named “unsure” if they didn’t feel confident about their rating.

I identify reverted edits by looking for subsequent revisions that completely discard the changes of a new editors' revisions using the approach for determining "identity reverts" described in chapter 2. Given Wikipedia's efficient vandal-fighting system, most unproductive edits are reverted within minutes, so 48 hours should be ample time to capture a reverting edit [? ]. We consider an editor to be "productive newcomer" if she makes at least one productive edit to an encyclopedia article within a week of her conversion.

To understand the cost of the increasing non-productive edits, we examined *how* these newcomer edits were reverted. The Wikipedia community's approaches to reverting damaging contributions can be organized into three categories:

- *manual* - Human editors revert edits directly via the web interface. Reverts performed manually require the most human effort.
- *semi-automated* - Wikipedians have developed a suite of tools to make the process of identifying and reverting damage require less human effort. Contributions revert by wikipedians using semi-automated tools require less effort than manual reverts.
- *automated* - Autonomous computer programs commonly referred to as "bots" monitor recent changes and perform reverts without human input. Contributions reverted by bots require essentially no human effort.

Advantageously, the *automated* and *semi-automated* tools identify themselves by leaving structured comments along with the revisions they make. Through the use of a set of regular expressions on these edit comments, we are able to categorize reverts into the above three categories. Reverts performed by user accounts with a bot "flag" were classified as *automated*. Reverts performed using Huggle<sup>8</sup>, Twinkle<sup>9</sup>, Popups<sup>10</sup>, Rollback<sup>11</sup> or STiki<sup>12</sup> were classified as *semi-automated*. All other reverts were classified as *manual*. We examined the proportion and raw number of reverts that are caught by each tool to understand the cost of non-productive newcomer contributions.

## 5.5 EXPERIMENTS AND RESULTS

### 5.5.1 RQ1: How do different elicitations affect the volume and utility of feedback?

To explore this research question, I randomly divided readers into three experimental conditions – one for each of the three feedback forms (1-3 from figure 25). For each condition, the appropriate feedback form appeared at the end of the encyclopedia article. Note that neither of the prominent links (labeled A and E) were used for this experiment.

<sup>8</sup> <http://enwp.org/WP:HG>

<sup>9</sup> <http://enwp.org/WP:TW>

<sup>10</sup> <http://enwp.org/WP:Popups>

<sup>11</sup> <http://enwp.org/WP:Rollback>

<sup>12</sup> <http://enwp.org/WP:STiki>

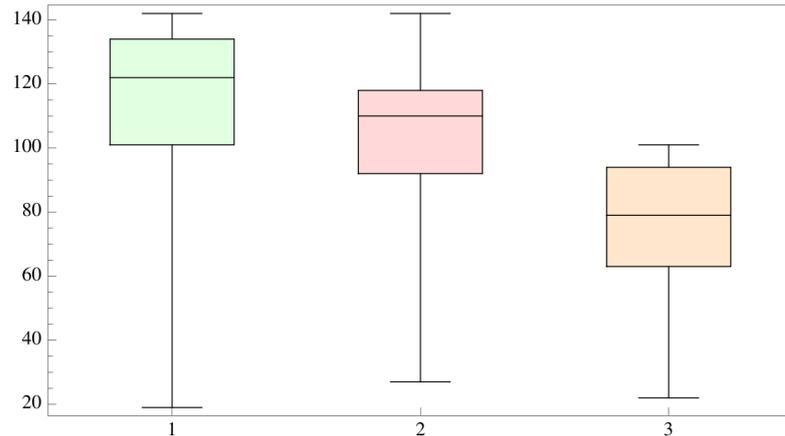


Figure 27: **Quantity of feedback by experiment.** The median daily feedback submissions per day is plotted for the last two weeks of the experiment with box limits at the 25% - 75% quantiles and bar limits at the most extreme observed values.

#### 5.5.1.1 Results

I considered different designs of the AFT as treatments applied to the same population of articles and measured the amount of daily feedback submissions generated in response of each treatment for the duration of the experiment (Dec. 27th - Jan 24th, 2012). Figure 27 plots the median submissions per day over the last two weeks of the experimental period. The first two weeks were discarded to limit the effect of novelty for the new feedback interface. Forms 1, 2 and 3 generated a total of 1666, 1539 and 1148 feedback submissions respectively. I used a one-way analysis of variance to test the null hypothesis that these treatments elicited an equal amount of feedback submissions per day. The test allows us to reject the null hypothesis at  $\alpha = 0.05$  ( $F(2, 42) = 6.57, p = 0.003$ ). Post-hoc comparisons using the Tukey range test indicates that the means for form 1 ( $M = 111, SD = 32.3$ ) and form 2 ( $M = 102.6, SD = 27.4$ ) were significantly higher than form 3 ( $M = 76.5, SD = 20.5$ ) at  $\alpha = 0.05$ . However, the rate of feedback submissions for form 1 did not significantly differ from form 2.

Next, I looked at the utility of feedback submissions via the three forms as determined by the Wikipedian evaluators. A random sample of up to 250 feedback submissions was gathered for each condition such that no two submissions came from the same article over the entire experimental period<sup>13</sup>. This sampling approach was used to produce statistics that reflect the expected utility of feedback submissions per article by controlling for the overrepresentation of popular articles. To measure the quality of feedback submitted, I examined the proportion of useful feedback submissions as determined via hand coding by multiple Wikipedians, as described in [Methods](#).

To combine the two ratings per feedback submission into a single assessment, I employed three aggregation strategies:

<sup>13</sup> Form 3 didn't elicit feedback submissions across enough distinct articles during this period, so I was only able to sample 143 submissions.

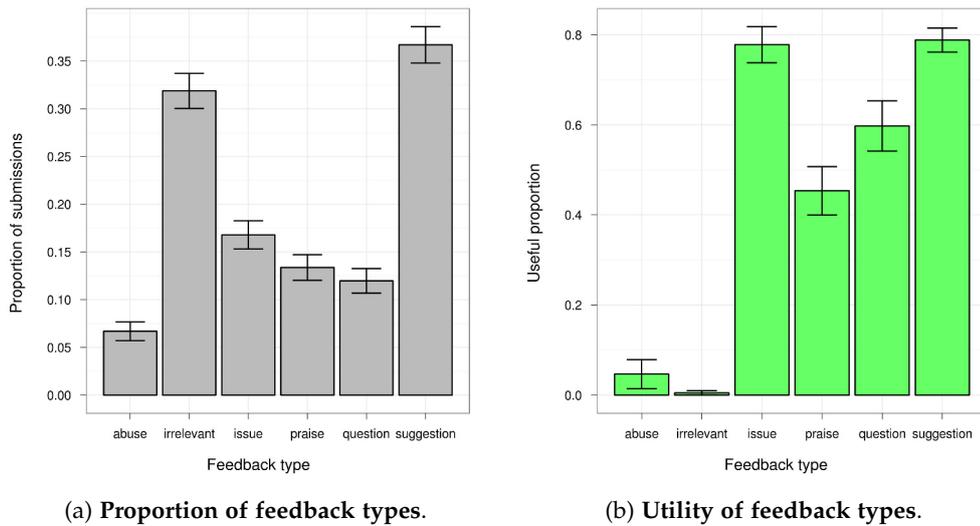


Figure 28: (left) The proportion of feedback submitted is plotted by intention as determined by at least one Wikipedian. (right) The proportion of useful feedback is plotted for each type. These plots draw aggregate proportions from feedback submitted through all three experimental conditions.

- **someone**: useful to at least one Wikipedian
- **both**: useful to both Wikipedians
- **strict**: useful to both Wikipedians and neither was unsure

The three rating aggregation strategies compare similarly between experimental conditions, so I opt to only report the proportion of feedback determined useful by both Wikipedians (“both” strategy).

The differences between the proportion of useful feedback submitted between the three interfaces varied insignificantly around 0.45. Of the three conditions, the most substantial difference was observed between form 2 (0.436) and 3 (0.469), however a  $\chi^2$  test found that difference to be insignificant ( $p = 0.604$ ).

I also found a relatively strong consistency between the apparent intentions of feedback submitted via the three interfaces with a couple of exceptions: a  $\chi^2$  test showed that form 2 elicited a significantly more useful *issues* than form 1 ( $\Delta = 0.225, p = 0.035$ ) and form 1 elicited a significantly higher proportion of questions than form 3 ( $\Delta = 0.100, p = 0.005$ ). Although these differences are statistically significant, the difference in useful issues could be considered insubstantial and may have come about by chance given the number of statistical tests performed to identify differences (6 types of feedback \* 2 tests \* 2 conditions = 24 tests).

To explore a potential mismatch between the types of feedback readers were likely to submit and the types of feedback Wikipedians found useful, I merged the feedback submitted via the three interfaces to compare the aggregate rate of feedback types and utility. Figures 28a and 28b show some substantial differences between the feedback submitted by readers and the feedback that Wikipedians found useful. As an example,

the proportion of feedback determined irrelevant by Wikipedians was the second most frequent type (31.9%) while the proportion of irrelevant feedback that was marked useful was vanishingly small (0.5%) and could be attributed to rater error given that it represents a single submission labeled both useful and irrelevant.

Figures 28a and 28b also shows some similarities. For example, suggestions appear to be both the most prevalent (36.7%) and useful (78.9%) submission type.

The relatively small proportion of abuse is also worth noting. Only 6.6% of feedback was determined to be submitted in bad-faith. To put this into context, 10.0% of anonymous edits to encyclopedia articles are explicitly labeled as vandalism when they are reverted.

#### 5.5.1.2 Discussion

Although the elicitation directed towards readers' concerns (form 1) did elicit more participation than the form directed toward editors' concerns (form 3), there was not a significant difference in the proportion of useful feedback. This result refutes **HYP Participation & utility** and suggests that the strategy of directing readers towards editor interests – at least in this case – did not affect the usefulness of their participation.

However, there was a slight difference to the types submitted feedback submitted. I observed a significantly lower proportion of questions submitted via form 3 which asks readers to rate an article. I suspect that, to evaluate the quality of an article, readers must put themselves in the role of an expert – someone qualified to perform an evaluation. Assuming readers who have legitimate questions would be less likely to adopt this role, it makes sense to see both a lower proportion of questions submitted and fewer overall submissions for form 3. Given the relatively high overall utility of questions (59.7% useful), this suggests that a substantial number of useful questions may not have been submitted by readers in this condition.

#### 5.5.2 RQ2: How does the prominence of the elicitation affect the volume and utility of feedback?

Given that the results reported in the last section suggest that the different forms had no substantial effect on the utility of submissions and the Wikipedian hand-coders preferred form 1, I continue experimentation using only that form. To look for changes in the utility of feedback submitted based on the amount of reader participation, I performed another experimental run using links "A" and "E" described in figure 25 in an attempt to increase the prominence of the feedback form. I re-shuffled readers into three new experimental conditions:

- **1X**: Form 1 loads only on the bottom of the article.
- **1A**: Same as 1X but includes link A
- **1E**: Same as 1X but includes link E

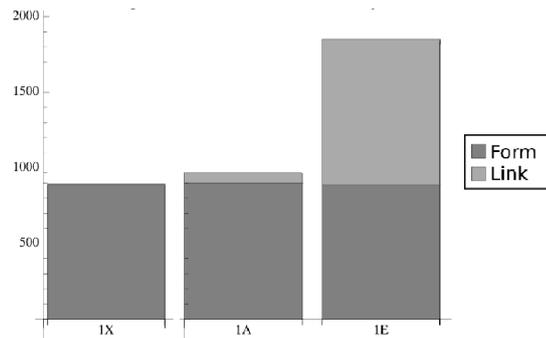


Figure 29: **Volume of feedback by experiment.** The raw amount of feedback submitted via the three experimental conditions is plotted by whether the feedback was submitted via the prominent link or via the form at end of articles.

### 5.5.2.1 Results

To ensure that the prominent links A and E were serving their intended function, I measured the total amount of feedback submitted over the period between April 5 and April 19 (15 days).

As figure 29 shows, the prominent link appears to have been effective in eliciting more feedback. While all cases elicited approximately the same amount of feedback directly via the form at the bottom of the article as the non-prominent condition (1X), the feedback submitted via the prominent link itself appears to purely supplement the total amount of feedback submitted (1X: 892, 1A: 967 and 1E: 1851). In the case of 1E, the prominent link apparently increased the rate at which feedback was submitted by about 91%.

To look for differences in the utility of feedback submitted via each of the experimental conditions, I randomly sampled 300 feedback submissions per condition and asked the Wikipedian coders to rate their usefulness. The differences between the proportion of useful feedback submitted between the three interfaces varied insignificantly around 0.42. A  $\chi^2$  of the largest observed difference between the conditions 1X (no link) and 1A (less prominent link) was insignificant ( $\Delta = 0.065, p = 0.124$ ). If the prominence of the interface caused a decline in the utility of feedback, I would expect to see a more substantial dip in the proportion of useful feedback in the 1E condition given the amount of additional feedback submitted.

However, I did observe a significant difference in utility based on the whether the feedback originated via the form at the bottom of the article or the prominent link E. Figure 30 plots the proportion of useful feedback submitted for each experimental interface by whether it originated via the form at the bottom of the article or the prominent link. The proportion of useful feedback submitted via the most prominent link (1E) was significantly lower than the feedback submitted via the corresponding form ( $\Delta = 0.153, p = 0.010$ ). This appears to be partially due to complaints about the prominent link submitted via the prominent link. After the experiment, I received messages from some concerned users that felt the prominent link was distracting.

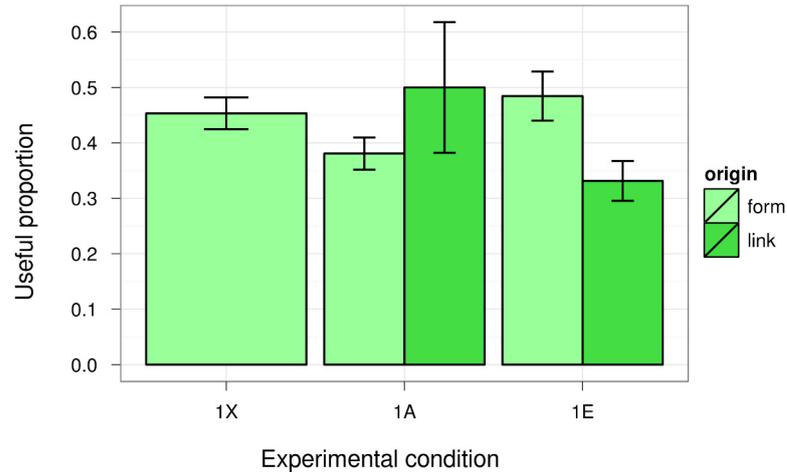


Figure 30: **Utility of feedback by origin.** The proportion of useful feedback is plotted for each condition by the origin from which it was submitted with standard error bars. Note that the large error bars around the proportion of useful feedback submitted via the link in the 1A condition is due to the small amount of feedback sampled in that condition ( $n = 18$ ).

Out of the 8 feedback comments submitted via 1E’s link by registered editors, only 1 feedback submission was determined to be useful while the rest were useless protests of the presence of the feedback interface.

#### 5.5.2.2 Discussion

Increasing the prominence of the feedback form increased the volume of contributions in the expected way, affirming **HYP Prominence & volume**. The results above suggest that the rate at which feedback is submitted can be nearly doubled by introducing prominent link E into the interface.

Surprisingly, I also found that expanding participation by making the interface more prominent did not affect the overall proportion of useful submissions, thus refuting **HYP Prominence & utility**. I hypothesized that increasing participation arbitrarily via a more obvious interface would elicit participation from readers who were increasingly less interested in contributing constructively, and therefore, utility would fall. However, I found no significant change in the proportion of useful feedback submitted despite more than doubling the amount of submissions. This result suggests that there are many potentially productive contributors available who may simply not be aware of their ability to contribute and that hiding the means to contribute at the bottom of the article is not an effective mechanism for improving the utility of contribution.

However, I did observe that feedback submitted via the most prominent link was of lower utility than that submitted directly via the form. Although it is tempting to conclude that the prominent link itself elicits lower utility submissions, it is important to note that the overall utility of submissions through the most prominent condition

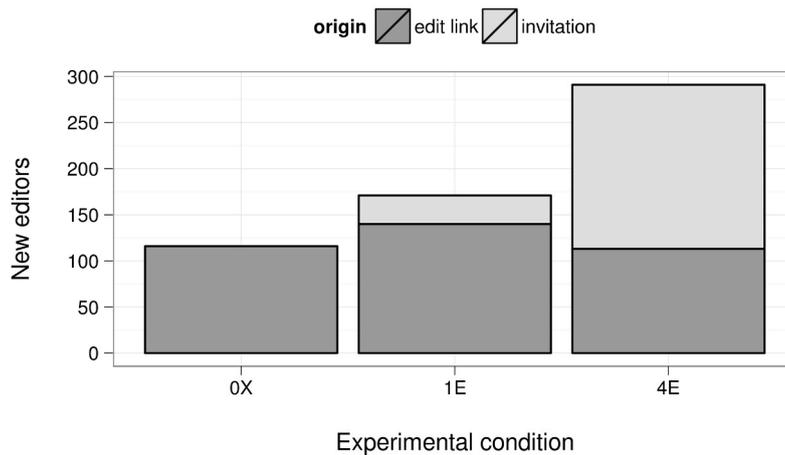


Figure 31: **New users by origin.** The number of new editors is plotted and stacked by the origin of their first edit. “edit link” refers to the standard vector for accessing the edit pane. “invitation” refers to form 4 which invites the user to make an edit.

(1E) was not significantly lower than the 1X condition despite the fact that the prominent link nearly doubled the rate at which feedback was submitted. Given the slightly larger proportion of useful feedback submitted directly via the 1E form than the 1X form, I suspect that the underlying cause of this disparity is a re-routing of less useful feedback (that would have been submitted anyway) through the prominent link.

### 5.5.3 RQ3: How does the presence of the feedback interface affect new editor conversion?

To look for a potential cannibalization effect of AFT on new editor conversions, I re-shuffled readers into three new experimental conditions:

- **0X:** Control condition (no feedback form or link)
- **1E:** Feedback form 1 is displayed at the bottom of the article, with prominent link E. Users were presented with edit invitation form 4 after successfully submitting feedback (indirect invitation)
- **4E:** Edit invitation form 4 is displayed at the bottom of the article, with prominent link E (direct invitation, no feedback form).

#### 5.5.3.1 Results

To measure new editor conversions, I observed user activity for the period between April 27th and May 7th 2012. During this observation period, both the direct invitation (4E) and indirect invitation (1E) conditions saw more new editors conversions than the control case. Figure 31 shows that a similar number of new editors originated via the standard vector (edit link) in the treatment conditions (1E:  $n = 140$ , 4E:  $n = 113$ ) as in the control condition (0X:  $n = 116$ ). The new editors whose first edit originated

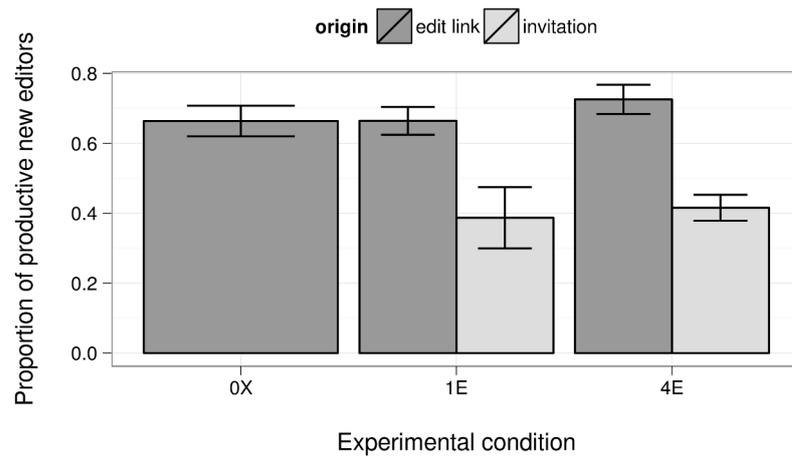


Figure 32: **Proportion of productive new users by origin.** The proportion of new editors who made at least one productive contribution in their first week is plotted by the origin of their first edit for the three experimental conditions.

via the invitations to edit (1E:  $n = 31$ , 4E:  $n = 178$ ) only appear to supplement these numbers.

To test if this increase in conversion was significant, I approximate the view-to-editor conversion rate by drawing 95% confidence intervals around the approximate views that each experimental condition received based on the total number of views that the sample received during the observation period (9,424,041). The random function which assigned readers to experimental conditions can be represented as a binomial proportion where the underlying probability of assignment to any one condition is  $\frac{1}{3}$ . Therefore, we can divide the total views and draw confidence intervals using the binomial approximation to a normal distribution ( $\frac{9424041}{3} = 3141347 \pm 945.47$ ). I used this confidence interval to perform a conservative  $\chi^2$  test that appropriately reduces the likelihood of a type I statistical error. Let  $u_1$  and  $u_2$  be set of new users such that  $|u_1| > |u_2|$  I performed the following test:

$$\chi^2 \left( \frac{|u_1|}{3141347+945.47}, \frac{|u_2|}{3141347-945.47} \right)$$

The conservative  $\chi^2$  test found the difference in the rate of new editor conversions to be significant for all three cases (0X < 1E:  $p = 0.002$ , 1E < 4E:  $p < 0.001$ ). This suggests that the indirect invitation condition (1E) converted significantly more views to new editors than control condition and the direct invitation condition (4E) converted significantly more views to new editors than both 1E and the control.

To explore the productivity of new editors conversions via the experimental conditions, I performed an analysis of the first week of contributions made by each new editor to look for productive contributions as described in section 5.4. I define a “productive editor” as a new editor who made at least one edit to an article that was not reverted within 48 hours by another editor.

As figure 32 shows, editors who originated via the invitation to edit through both the direct and indirect call to action were significantly less likely to make produc-

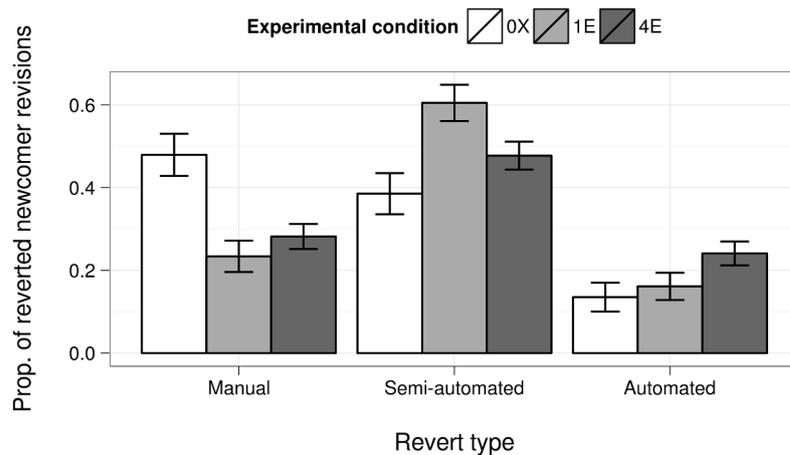


Figure 33: **The cost of non-productivity.** The proportion of newcomer revisions reverted for each experimental condition by how the revision was reverted.

tive contributions in their first week than those who originated through the edit links within the same condition (1E:  $\Delta = 0.277, p = 0.008$ ; 4E:  $\Delta = 0.310, p < 0.001$ ). When comparing only those new editors who originated via the edit link between experimental conditions, the proportion of productive new editors varied insignificantly around 0.68. These results suggest that new editors who originated via an invitation to edit are about half as likely to contribute productively in their first week as those editors who originated via the standard vectors.

To examine the cost of introducing more unproductive newcomer contributions to Wikipedia, I classified the reverted newcomer editors in each condition into *manual*, *semi-automated* and *automated* as described in section 5.4. As figure 33 suggests, a  $\chi^2$  test found that a significantly smaller proportion of unproductive newcomer edits were reverted manually in both the 1E ( $\Delta = -0.25, p < 0.001$ ), and 4E ( $\Delta = -0.19, p = 0.001$ ) conditions. A significantly larger proportion of 1E revisions were reverted with semi-automated tools ( $\Delta = 0.219, p = 0.002$ ) and that a significantly larger proportion of 4E revisions were reverted with fully automated tools ( $\Delta = 0.134, p = 0.019$ ) when compared to the control. These differences suggest that Wikipedia’s damage filtering systems (automated and semi-automated tools) are making up for some of the increased cost of damage due to the invitation to edit.

### 5.5.3.2 Discussion

The results described in this section are a solid refutation of **HYP Conversion**. Even without including those editors who saw the edit invitation, (insignificantly) more new editors made their first contribution in the 1E condition through the standard vectors to the edit pane (edit link) than in the control case. The indirect call to action appeared to only supplement new editor conversions. Further, this class of new editors appears

to be just as likely to be productive in their first week of tenure as those who started editing via the control case.

However, there appears to be a cost to boosting conversions via invitations to edit – at least initially. A substantially smaller proportion of editors originating via the invitation made a productive contribution in their first week of edit activity.

As I suspected, the edit invitation strategy is a double-edged sword. Although the invitation appears to have effectively convinced substantially more readers to try their hand at editing, these invited editors were less likely to make productive contributions in their early wiki-career than those who started editing on their own. This result supports **HYP *Converted productivity***.

Social learning theory suggests that newcomers to a community tend to go through a natural process by which they learn about a community and build confidence before expanding their contribution pattern[73]. It seems likely that explicitly inviting new editors to contribute would artificially inflate their confidence and speed up the process of introduction to the community which would naturally lead to more mistakes – at least initially.

My analysis suggests that, at least in part, Wikipedia’s efficient damage filtering mechanisms reduce the cost of the increase in unproductive newcomer contributions. Although I can assume that the cost in human effort of a few more automated reverts is essentially zero, the cost of operating the semi-automated reverting tools is unclear. Without such understanding, it’s difficult to reason about just how much Wikipedia’s damage filtering mechanisms make up for the increase in unproductive newcomer contributions.

## 5.6 CONCLUSIONS

In this study, I examined a strategy for increasing participation from Wikipedia’s consumers (readers) through a set of experiments performed using the Article Feedback Tool, an extension of the software that allows encyclopedia readers to make simple and low risk, yet productive, contributions – thereby supporting legitimate peripheral participation. Table 8 summarizes the results of our three experiments in the context of the research questions and hypotheses stated in section 5.3.

**Quantity vs. quality of participation.** Although I hypothesized a trade-off between the quality and quantity of participation, my observations of such an effect were inconsistent. While I was able to boost the rate of feedback submissions both by aligning AFT’s feedback forms to readers concerns and by making the interface more prominent to readers, I did not observe a decrease in the quality (“useful”ness as determined by Wikipedian raters) of feedback submitted. This result reflects the findings of Wash and Lampe[95] and suggests that there are many potentially productive non-contributors available who may simply be unaware of their ability to contribute and that obfuscating the mechanism for contribution is not an effective mechanism for improving the quality of contribution. In other words, increasing transactional costs

RQ1: How do different elicitations affect the volume and utility of feedback?		
<b>HYP Participation &amp; utility</b>	unsupported	Although form 1 and form 2 elicited more feedback submissions than form 3, the quality of submissions was insignificantly different.
RQ2: How does the prominence of the elicitation affect the volume and utility of feedback?		
<b>HYP Prominence &amp; volume</b>	supported	The prominent link conditions elicited significantly more feedback than the control.
<b>HYP Prominence &amp; utility</b>	unsupported	Although feedback submitted via the prominent links was less useful, the overall differences were insignificant.
RQ3: How does the presence of the feedback interface affect new editor conversion?		
<b>HYP Conversion</b>	unsupported	The presence or absence of the feedback interface had no observed effect on the rate of new editor conversions through the standard vectors.
<b>HYP Converted productivity</b>	supported	New editors whose first edits originated via an invitation were substantially less likely to contribute productively in their first week.

Table 8: A summary of hypotheses and findings.

either through requiring the reader to occupy a certain role or through obfuscating the ability to contribute do not appear to be effective strategies for increasing contribution quality for this low investment type of contribution. On the contrary, both our results and those of Wash and Lampe [95] suggest that participation can be increased by making the means of contribution more prominent without sacrificing quality with simple and straightforward contribution types like comments and feedback.

However, when AFT invited editors to edit articles, those supplemental new editors were less likely to be productive in their first week of tenure in Wikipedia. Editing an encyclopedia is a more demanding type of contribution than leaving a comment or submitting feedback, because editors must both contribute novel encyclopedic information and work within Wikipedia’s complicated set of policies and guidelines – a requirement that new editors tend to struggle to meet (as we saw in chapter 4). Under the framing of legitimate peripheral participation, we might expect potential new editors to naturally go through a lurking process by which they build a situated understanding of the Wikipedia community and their place within it before making their first contribution – an assertion supported by [3]. When viewed this way, the invitation to contribute could be short circuiting the natural “joining script” [92] of new Wikipedia editors. If this is true, other systems with non-trivial contribution difficulty should see a similar effect.

**System utility.** It is useful to understand the value of AFT’s invitation to new editors to the Wikipedia community’s goals from a system-level perspective. In order to

draw a data-based recommendation on whether AFT's invitation to edit should be released at a larger scale or not, both the value of new productive contributions and the cost of unproductive contributions must be accounted for. Systems with a high value on new contributions/contributors and/or efficient filters for dealing with unwanted contributions will be more likely to benefit from invitations despite the increase in unproductive edits.

Related work by Geiger et al. suggests that the English Wikipedia's automated and semi-automated anti-vandal tools are an efficient and scalable solution to the problem of moderation [30]. Our analysis suggests that these tools are effectively reducing the increased cost of unproductive contributions via the invitation to edit. In this context, it seems likely that the invitation to edit represents a net benefit to Wikipedia despite the larger proportion of unproductive edits. Related works examining Slashdot's distributed moderation system [53] and YouTube's copyright infringement detection algorithms [86] suggest a trend toward efficient, distributed, and automated curation mechanisms in Web 2.0 systems that may also reduce the cost of such unwanted contributions in other systems.

# 6 SNUGGLE

---

## 6.1 INTRODUCTION

In 2006, when both the number of volunteer editors and the size of the encyclopedia was growing exponentially, Wikipedia’s “vandal fighters” came to see “the free encyclopedia that anyone can edit” as a firehose of changes needing constant surveillance. To make this practical, they formalized the practice of reviewing edits around a suite of semi-automated tools (discussed in chapter 4). The design of these tools reflected these Wikipedian’s vision of Wikipedia.

As a result, today’s vandal fighters see a much different Wikipedia than most people do. In one sense, this is a metaphorical statement about “social worlds” [81], where people in certain cultures or organizations learn such different frameworks for interpreting their experiences that they can be said to inhabit different worlds. Yet in another sense, these vandal fighters *literally* see something different when they look at Wikipedia, since their work often consists not of browsing to Wikipedia pages in a web browser, but rather using a standalone desktop application called Huggle. This power tool presents vandal fighters with a queue of edits to review. With one click, they can instantly reject an edit and send its author a pre-written warning.

Huggle and related tools were developed by and for Wikipedians to them perform critical quality control tasks that are still necessary to this day [28]. These tools have become the standard way in which newcomers are “welcomed” to Wikipedia [32]. However, these tools raise practical design challenges and ethical issues for HCI researchers. The tools embody particular assumptions and values: they situate users as police, not mentors, who reject and punish, rather than encourage and support. We saw the results of these design choices in chapter 4; Wikipedia’s ability to recruit and retain new editors has been compromised. As Wikipedia is a complex socio-technical system, both its problems and solutions are likely to be just as complex; so how should we approach designing a solution?

In contemporary HCI, there are two general approaches to tackling these kinds of problems (c.f. [36]). “Second-wave” HCI seeks to design efficient interfaces and systems to support desirable social tasks, such as a tool to support mentors and the productive socialization work they (could) do. Thinking like a second-wave HCI designer, success is based on building an interface that affords the kind of capabilities that users require. Under this way of thinging, Wikipedia will become a better place for newcomers if enough people use a properly-designed socialization tool. “Third-wave” HCI is more *critical* in its approach – seeking targeted design interventions that interrogate the ideological foundations undergirding practices. Thinking like a third-

wave HCI designer, if the fundamental assumptions that gave rise to Huggle are not questioned, then a new tool may only be a temporary fix to a larger systemic problem.

Much ink and bits have been spent debating the relative merits of these two approaches to HCI, but as I pursued a solution, I found myself incorporating both second- and third-wave ways of thinking about designing and evaluating software systems, and further, that these perspectives were mutually supporting.

In this chapter, I'll describe the design and evaluation of *Snuggle*, a collaboratively-designed newcomer socialization system. This work contributes to HCI in three ways, by presenting Snuggle as (1) an instance of an intelligent newcomer socialization tool that increases task efficiency and supports the development of new norms of practice, with implications for other open systems; (2) as an account of a "successor system", a way of doing ideological critique through the design of systems that build better accounts of the world, integrating approaches to HCI which are often cast as incommensurable; and (3) a case of a highly-participatory design process, in which situated methods were not only shape the design of the interface, but also the design of the design process itself.

## 6.2 WIKIPEDIA'S SOCIO-TECHNICAL PROBLEMS

In 2006, when both the number of volunteer editors and the size of the encyclopedia was growing exponentially, Wikipedia's "vandal fighters" came to see "the free encyclopedia that anyone can edit" as a firehose of changes needing constant surveillance. To make this practical, they formalized the practice of reviewing edits around a suite of semi-automated tools (discussed in chapter 4). The design of these tools reflected these Wikipedian's vision of Wikipedia.

As a result, today's vandal fighters see a much different Wikipedia than most people do. In one sense, this is a metaphorical statement about "social worlds" [81], where people in certain cultures or organizations learn such different frameworks for interpreting their experiences that they can be said to inhabit different worlds. Yet in another sense, these vandal fighters *literally* see something different when they look at Wikipedia, since their work often consists not of browsing to Wikipedia pages in a web browser, but rather using a standalone desktop application called Huggle. This power tool presents vandal fighters with a queue of edits to review. With one click, they can instantly reject an edit and send its author a pre-written warning.

Huggle and related tools were developed by and for Wikipedians to them perform critical quality control tasks that are still necessary to this day [28]. These tools have become the standard way in which newcomers are "welcomed" to Wikipedia [32]. However, these tools raise practical design challenges and ethical issues for HCI researchers. The tools embody particular assumptions and values: they situate users as police, not mentors, who reject and punish, rather than encourage and support. We saw the results of these design choices in chapter 4; Wikipedia's ability to recruit and retain new editors has been compromised. As Wikipedia is a complex socio-technical

system, both its problems and solutions are likely to be just as complex; so how should we approach designing a solution?

In contemporary HCI, there are two general approaches to tackling these kinds of problems (c.f. [36]). “Second-wave” HCI seeks to design efficient interfaces and systems to support desirable social tasks, such as a tool to support mentors and the productive socialization work they (could) do. Thinking like a second-wave HCI designer, success is based on building an interface that affords the kind of capabilities that users require. Under this way of thinging, Wikipedia will become a better place for newcomers if enough people use a properly-designed socialization tool. “Third-wave” HCI is more *critical* in its approach – seeking targeted design interventions that interrogate the ideological foundations undergirding practices. Thinking like a third-wave HCI designer, if the fundamental assumptions that gave rise to Huggle are not questioned, then a new tool may only be a temporary fix to a larger systemic problem.

Much ink and bits have been spent debating the relative merits of these two approaches to HCI, but as I pursued a solution, I found myself incorporating both second- and third-wave ways of thinking about designing and evaluating software systems, and further, that these perspectives were mutually supporting.

In this chapter, I’ll describe the design and evaluation of *Snuggle*, a collaboratively-designed newcomer socialization system. This work contributes to HCI in three ways, by presenting Snuggle as (1) an instance of an intelligent newcomer socialization tool that increases task efficiency and supports the development of new norms of practice, with implications for other open systems; (2) as an account of a “successor system”, a way of doing ideological critique through the design of systems that build better accounts of the world, integrating approaches to HCI which are often cast as incommensurable; and (3) a case of a highly-participatory design process, in which situated methods were not only shape the design of the interface, but also the design of the design process itself.

### 6.3 DESIGN STRATEGY

When I began considering potential solutions to Wikipedia’s socialization problem, my goal was simple: I wanted to figure out how to solve the problem. I wasn’t looking for an opportunity to design a user interface or critique HCI. The idea for building a new user interface was born from this goal because it seemed like the most effective way to solve the problem.

As I discussed in section 6.1, my initial formulation of Snuggle was based on my concern with how the existing counter-vandalism tool, Huggle, framed newcomers’ activities as problems to be dealt with. My previous work suggests that the widespread use of this lens of suspiciousness has immediate implications: newcomers who do not already know all of Wikipedia’s rules (and how would they!?) are “bitten”<sup>1</sup> and

<sup>1</sup> “Bite” in Wikipedian jargon means aggressive actions towards newcomers. See <http://enwp.org/WP:BITE>

will leave the project before mentors can come help. However, this also has systemic, long-term implications in that viewing newcomers through lenses of quality control and counter-vandalism situates newcomers as inherently suspicious, rather than people who may make well-intentioned mistakes while learning how to be a part the community.

I knew of many Wikipedians who were interested in mentoring and socialization, so I saw an opportunity to design a tool that would support their practices just as Huggle supports vandal fighters. However, if vandal fighters continued their gate-keeping unabated, Snuggle might be only stop-gap attempt to retroactively respond to newcomers who already had been bitten. To further complicate the picture, Huggle serves a real need: vandalism was and is real, and is a real threat to Wikipedia. In fact, some of my collaborators and beta testers were also users of “competing” tools, and emphasized the need for Snuggle to support tasks like requesting an admin block a problematic newcomer. I empathize with these needs, and I do not see vandal fighters as “the enemy” and Snuggle as a tool to equip an army of “vandal fighter fighters”.

So how can I hope to solve the underlying problem? It became clear to me early on *I* couldn’t single-handedly solve any problems. What I somehow needed to accomplish with Snuggle was to change the way that Wikipedians thought about newcomers so that they could re-construct their practices such that good newcomers could be supported *and* efficient quality control could still preserve the content of the encyclopedia. This would require starting discussions, encouraging critique, and changing minds – both about how newcomers are currently treated and about how newcomers ought to be treated.

I was inspired by third-wave HCI research that critique dominant ideologies with the goal of reversing the assumptions that create complex social problems [38, 16]. In reflecting on my role as both researcher and Wikipedian, I saw the potential for Snuggle as part of a broader conversation about participation, representation, and inclusion in Wikipedia. Snuggle could provide the visibility to add an essential thread to the ongoing conversation about righting Wikipedia’s discontents; this conversation already includes discourse (e.g. essays<sup>2</sup>, tools (e.g. wikilove<sup>3</sup> and The Wikipedia Adventure<sup>4</sup>), and spaces (e.g. the Teahouse[64] and the Adopt-a-user program).

Throughout my work studying the Wikipedia community, I’d inadvertently become a community member myself. This proved to be an essential asset when considering how I could design Snuggle to effect the change I saw as necessary from a purely quantitative point of view. As an HCI researcher situated in Wikipedia, I didn’t just have a better sense about how to support mentoring practices; I also had a better sense about what kinds of conversations were taking place and what was missing from them. I sought to give all Wikipedians – not just dedicated mentors – a tool for finding, exploring, and reflecting on cases where newcomers were making good faith efforts to contribute, but made mistakes that were seen as vandalism. Thus, I

<sup>2</sup> Wikipedians write essays as a sort a descriptive and critical practice[62] and op-eds

<sup>3</sup> A tool designed support the practice of acknowledging other Wikipedia editors of their good work.

<sup>4</sup> A video game-like socialization tool that walks newcomers through a curriculum.

intended for Snuggle to serve to both (a) support early and positive mentoring and (b) make underlying assumptions and practices visible to Wikipedians to enable reflection and critique. Put crudely, Snuggle is intended to improve newcomer retention both directly (by enabling support of bitten newcomers) and systemically (by changing how experienced Wikipedians view newcomers).

#### 6.4 HOW DO WE EFFECT CHANGE?

How does the HCI literature suggest I design Snuggle? In this section, I'll cover literature from both the second- and third-wave perspectives that I

I wanted Snuggle to be an effective tool for Wikipedians who wanted to find and help newcomers in need. Classic 'second-wave' user-centered design literature was helpful to this end. Harrison et al. describes this work as asking the question: "What would we do differently because of the observations or findings that come out of these approaches?" [36] User-centered design emphasizes three factors: iterative design, empirical measurement, and a focus on users and tasks [? ]. Later research emphasized the co-evolution of systems and practices based on the insight that introducing a new tool transforms the user's tasks and context [70]. Frameworks such as activity theory [67] and distributed cognition [37] take into account how action or cognition is situated in a diverse set of technological and social contexts. Design approaches such as ethnographically-informed design [? ] and participatory design view people not as users to be designed *for*, but as people to design *with* [65].

I was also inspired by approaches from third-wave HCI; their ideological critiques of dominant systems resonated with the problems I saw in vandal fighting tools. I also saw similarities with the "values in design" literature, where designers explicitly acknowledge principles they value and seek to uphold [22]. I also saw alignment with "action research" [? ], which aims to bring about large-scale social, cultural, political, economic, or environmental benefits through research work.

Yet much of third-wave HCI work has been framed as inherently oppositional to second-wave HCI [36]. However, I do not see it this way. Instead, I find Bardzell and Bardzell's (bardzell11towards) perspective on feminist HCI more congenial. They argue that feminism is neither qualitative nor quantitative, that HCI is not inherently antifeminist, and raise the possibility of "integrating feminist epistemologies with the epistemologies behind much of the best quantitative work in HCI." They leave this as an open challenge, which I accept.

##### 6.4.1 *Successor systems*

In thinking about the Snuggle design process and related work in the HCI literature, Haraway's notion of "successor science" [35] resonated. She called for new forms of science that blend objectivity with situatedness: "Feminists have to insist on a better account of the world; it is not enough to show radical historical contingency and modes

of construction for everything.” This work is based on criticism of how newcomers in Wikipedia are often seen through one particular lens representing the vandal fighter’s perspective, and in designing Snuggle, I attempted to create a better account of newcomers.

This concept of a successor system<sup>5</sup> – doing ideological critique by designing and deploying systems that help build better accounts of the world – was helpful in situating Snuggle in relation to other related HCI projects. Hollaback represents a critique the widespread institutional ignorance of street harassment in two related ways: it first provides a safe space for victims of street harassment to assemble as a networked public, and second provides an infrastructure for building better accounts of the world, ones that make often-ignored experiences of street harassment visible at a variety of scales [16]. Turkopticon embodies a critique of the way Amazon’s Mechanical Turk turns human workers into an invisible, de-individuated infrastructure, ripe for exploitation with little to no recourse [38]. As design activism, Turkopticon affords workers the ability to rate employers, building a better account of the world for two purposes: to “not only hold employers accountable, but induce better behavior.”

## 6.5 DESIGN OF SNUGGLE

### 6.5.1 *Participatory design process*

Wikipedians have built their own infrastructure and processes for creating new tools, extensions, and bots, which I used to design Snuggle. Participatory design took place in routine spaces in Wikipedia. I maintained an evolving wiki page where I described the project, published prototypes, and recruited collaborators and testers. I used standard wiki talk pages as a forum throughout the design process, bringing individual concerns and conversations about the design to a wider audience. I also added new design elements as probes to provoke discussion about mentoring norms (see section 6.5.5 below). The talk page was active and successful, with 23 distinct editors who sent 107 messages.

My Wikipedian collaborators influenced not only the design of Snuggle, but also the design of the design process by helping me aligning my approach with Wikipedia’s norms. Practices regarding releasing updates and changelogs had to be mutually negotiated. My collaborators also worked to recruit other Wikipedians, facilitated discussions, and created some of the spaces in which I did participatory design. One even created an IRC channel for Snuggle users and configured a bot to post messages to the channel when I posted a new update or design.

---

<sup>5</sup> My close collaborator, Stuart Geiger, deserves credit for the initial formulation of the “successor systems” idea that we refined together for this work.

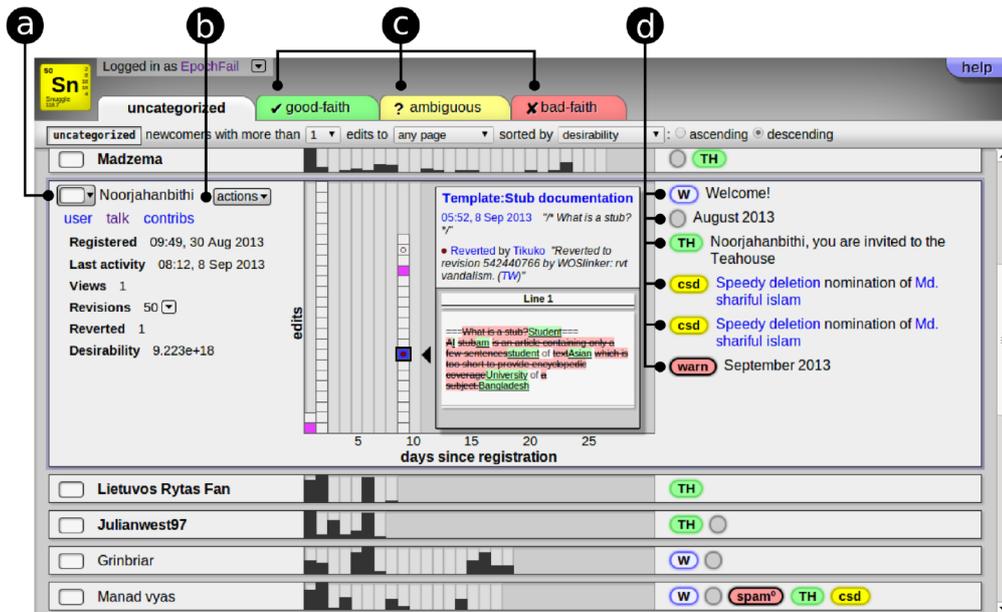


Figure 34: **Snuggle’s user browser.** A screenshot of the Snuggle user browser is presented with UI elements called out. The user dossier for “Noorjahanbithi” is selected. An edit in the *interactive graph* is selected and information about the edit is presented. (a) The (unexpanded) categorization menu, (b) The (unexpanded) wiki action menu (see section 6.5.4.2), (c) Tabs for accessing lists of categorized users, and (d) Talk page icons representing socially relevant traces (see section 6.5.4.1).

### 6.5.2 System overview

Snuggle tracks recent activity in the English Wikipedia by reading “recent changes”<sup>6</sup> from the website’s API<sup>7</sup>. Using this feed of activity, Snuggle builds *user dossiers* on all newcomers who have registered within the last 30 days. User dossiers include activity statistics, an interactive graph of edits, and a visual summary of interactions they’ve had with other editors extracted from their talk pages (see section 6.5.4).

The interface displays user dossiers in four lists: uncategorized, good-faith, ambiguous, and bad-faith (figure 34c). Snuggle users move dossiers between the four lists by using the categorization menu (figure 34a), and perform relevant tasks in Wikipedia using the wiki actions menu (figures 34b and 36).

### 6.5.3 Desirability sorting

Every day, about 1,000 people register an account and make at least one edit to English Wikipedia. Mentors can’t wade through that many newcomers unless they devote several hours a day to the work and abandon encyclopedia writing entirely. Snuggle needed to efficiently support identifying desirable newcomers.

<sup>6</sup> a chronologically sorted list of most activities that take place on Wikipedia from edits to new user registrations

<sup>7</sup> <http://en.wikipedia.org/w/api.php>

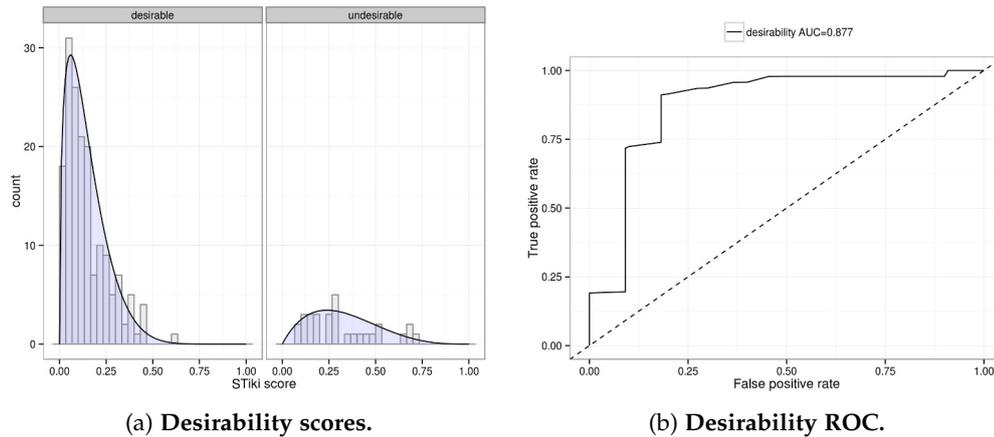


Figure 35: (left) Histograms of the frequency of STiki scores are plotted for the training set newcomers' edits with expectation maximization fits of beta distributions overlaid. (right) The receiver operating characteristic of the desirability ratio of newcomers from the test set is plotted.

**Desirability in concept.** Wikipedians refer to the desirable behavior of others as “good-faith”. In Wikipedia, the concept of “good-faith” is based on the intention of a user as opposed to the effects of their actions. When discussing newcomers, the Assume Good Faith guideline states<sup>8</sup>:

A newcomer’s behavior probably seems appropriate to him/her, and a problem in that regard usually indicates unawareness or misunderstanding of Wikipedia culture.

The guidelines stresses the importance of seeing damaging edits as mistakes rather than as intentional.

**Desirability in practice.** Classifying the intentions of newcomers as good-faith or bad-faith is a necessary part of interacting with others in Wikipedia. Because of Wikipedia’s openness and popularity, it’s very common to find newcomers who are purposefully trying to make offense or otherwise waste time. To help mentors efficiently prioritize *desirable* newcomers to assist, I sought to rank newcomers using a similar modeling technique as counter-vandalism tools, but with the opposite valence. In order for this model to be useful in Snuggle, the modeling strategy needed to:

- Make useful judgments about newcomers who have made few edits.
- Refine these judgments as new information (edits) becomes available.
- Not be based on the negative reactions received by newcomers (e.g. reverts, warnings, and blocks) since those are useful cues for identifying mentorship opportunities.

<sup>8</sup> <http://enwp.org/WP:AGF>

**Information source.** One approach is to sort newcomers by the proportion of their edits that have been reverted, but this defeats the broader goals of Snuggle. If only newcomers who are least reverted are determined to be working in good-faith, then Snuggle would not be a useful tool for identifying good-faith newcomers who are reverted due to mistakes or misunderstandings.

In order to avoid considering vandal fighters’ reactions to newcomers, I strategically take advantage of sophisticated models already used to assess newcomer behavior in Wikipedia: counter-vandal bots. Many of these bots publish scores of individual edits, based on the probability that the edit is vandalism. I suspected such scores would be useful for differentiating the activities of good-faith newcomers from bad-faith newcomers, independent of whether or not the edits were eventually reverted.

**Modeling desirability.** I constructed a Bayesian model by intersecting a dataset of newcomers hand-coded as “desirable” and “undesirable” from chapter 4 with scores retrieved from STiki’s API<sup>9</sup> to arrive at 152 hand-coded newcomers and 377 scored “first session”<sup>10</sup> edits.

I randomly split the set of newcomers in half to create a pair of training and test sets (76 users/set). I then used an expectation maximization approach to fit two beta distributions to the training set scores for desirable and undesirable users. Using these two distributions as models for STiki scores attributable to desirable and undesirable editors (see figure 35a), I use the following function to generate the odds than an editor is editing in good faith given a set of STiki scored edits:

$$\text{desirability ratio} = \frac{p(\text{scores}|\text{desirable})p(\text{desirable})}{p(\text{scores}|\text{undesirable})p(\text{undesirable})} \quad (4)$$

With this approach, I was able to attain a relatively high AUC (0.877) using scores from edits that newcomers performed in their first session.

#### 6.5.4 *Social literacy via traces*

To visualize socially relevant activity and act in Wikipedia, Snuggle took into account the structured documentary traces Wikipedians developed to coordinate and share information. These traces are a core component of social and organizational interaction in not just Wikipedia, but a variety of “virtual” and traditional co-located organizations. These traces “not only document events but are also used by participants themselves to coordinate and render accountable many activities.” [31] For example, vandal fighters use traces to track newcomers, using their public talk pages as a kind of shared database to determine how close they are to being blocked from editing [30]. Understanding traces is part of what it means to be a Wikipedian, and traces are fol-

<sup>9</sup> see <http://enwp.org/WP:STiki> and [97]

<sup>10</sup> An edit session is concept formalized in [29] that temporally clusters edits together into “sessions”. A users “first session” represents their first editing experience as a registered editor.

lowed and left in performing many socially relevant actions. Ford & Geiger argue that newcomers who are not yet “trace literate” suffer from power imbalances by being unable to participate effectively or know how they are being tracked [23].

In order for Snuggle users to take into account the actions taken against newcomers – as well as to have their own actions affect Wikipedia – Snuggle will have to consume and produce traces.

#### 6.5.4.1 Trace consumption

For Snuggle’s trace consumption, I focused on newcomers’ talk pages. As mentioned previously, a user’s talk page is used both to capture one-on-one conversations as well as to document the interactions that the user has had with Wikipedia’s quality control system.



Figure 36: **Wiki actions menu.** A screenshot of the “wiki action menu” is presented with a the message sending functionality selected and a test message written. Note the preview on the right side specifies which page the message will be appended to.

Luckily, the structure of these traces is highly consistent because Wikipedians mostly use templated messages when interacting with newcomers. This consistency lends itself to detectability, so I was able to define whole classes of traces with a simple set of regular expressions. Snuggle represents traces with icons on the right side of the user dossier (see figure 34d).

#### 6.5.4.2 Trace production

In order to support Wikipedians’ work practices, some actions will need to be performed back in the wiki. Snuggle supports relevant newcomer mentorship and routing actions: send a message, invite to the Teahouse<sup>11</sup> and report vandalism<sup>12</sup>.

In Wikipedia most traces are preserved via edits to pages, so I developed a configurable trace production system capable of previewing and producing page edits and other actions that these traces represent. Figure 36 shows the “wiki actions menu” with a form describing the action to be performed on the left and a preview of resulting page edits on the right.

<sup>11</sup> a newcomer socialization space discussed in [64]

<sup>12</sup> posts to a forum for bringing bad-faith editors to the attention of administrators

6.5.5 *Social translucence*

While my design process gave me insight into what tasks Snuggle should support, there's a whole separate set of work practices that must be built around the system by its users. In order to support the *social processes* surrounding the development of these new practices, I took inspiration from Erickson & Kellogg's work describing the design of socially translucent systems [20]. They argue for three characteristics of social translucence: visibility, awareness, and accountability.

Snuggle makes the activities of Snuggle users both visible and prominent. I designed a screen to welcome for then users first load Snuggle that displays a list of recent activities performed by Snuggle users (see figure 37). This recent activity list doesn't only support visibility, but the ability to react as well. Clicking on the username of the newcomer acted upon will open the user dossier for that newcomer complete with categorizer and wiki action menu. Most critically, this transparency is made apparent to Snuggle users before they've had an opportunity to log in and begin using Snuggle.

Recent activity	
all events ▾	by user <input type="text"/>
09:45, 7 Sep 2013	Chris troutman invited <input type="text" value="How Shuan Shi"/> to the teahouse
05:25, 7 Sep 2013	Buster7 categorized <input type="text" value="Ncnative556"/> as <span style="background-color: #d4edda;">good-faith</span> <i>Maybe an SPA</i>
05:13, 7 Sep 2013	Buster7 invited <input type="text" value="Mostly home"/> to the teahouse
05:06, 7 Sep 2013	Buster7 invited <input type="text" value="Jmlow"/> to the teahouse
05:06, 7 Sep 2013	Buster7 categorized <input type="text" value="Jmlow"/> as <span style="background-color: #d4edda;">good-faith</span> <i>(no comment)</i>
05:01, 7 Sep 2013	Buster7 invited <input type="text" value="Blue0burak"/> to the teahouse

Figure 37: **The recent activity feed.** A screenshot of Snuggle's recent activity list is presented.

This visibility serves some practical functions as well. Allowing Snuggle users to observe each others activity may help allay concerns about bad behavior by encouraging feelings of accountability. Snuggle's users may think carefully about their actions when they know they are visible to other Snuggle users and by enabling Snuggle users to engage in peer-policing they should be able to detect and address troublesome users quickly.

## 6.6 WHAT IS SNUGGLE?

### 6.6.1 *Snuggle as a newcomer socialization tool*

Snuggle is designed to support a specific newcomer socialization task, the detection of newcomers in need of help. In a way, this manifests similarly to anomaly detection. Snuggle's sorting based on desirability ratio is designed to bring attention to good-faith newcomers and the trace consumption system is designed to bring attention the negative attention that these good-faith newcomers receive. In a perfect system, good-faith newcomers would not receive such negative attention.

Support for the detection of these anomalous situations is intended to make the process of identifying desirable newcomers who need support faster and more effective than the passive model currently employed by Wikipedians. Snuggle users can use the talk page trace visualization to identify warnings, deletion notifications, and other negative reactions that newcomers received and compare this response with the edits presented in the interactive graph in order to target newcomers in need of support. In other words, by juxtaposing a desirable newcomer's activities with the negative reactions they received, I hope to enable mentors to selectively pay attention to newcomers most likely to benefit from it.

By providing a means for mentors to selectively intervene when desirable newcomers are treated badly, Snuggle enables Wikipedian mentors to more effectively deal with the immediate concerns raised in chapter 4.

### 6.6.2 *Snuggle as a critique/successor system*

Snuggle also serves as a critique of the way that counter-vandalism tools represent newcomers and the effects that has had on the efficiency of Wikipedia. In the design of Snuggle, I aim my critique at three characteristics of counter-vandalism tools: edits as the unit of assessment, sorting by undesirable only, and primarily supporting negative reactions.

**Edits as the unit of assessment.** Counter-vandalism tools like Huggle only show their users a single edit at a time. An individual edit, taken out of context, is a very limited frame by which to view a newcomer's activities. Given this limited amount of information, I suspect that it is much easier to ascribe simple *good* or *bad* to complicated individuals. In other words, it's easy confuse mistakes or even good work that just happens to look suspicious<sup>13</sup> as the work of vandals (false positives).

I sought to provide Snuggle's users with a more complete view of the newcomers they interact with. The first component of Snuggle that I designed was the user dossier that brings together as much information as I could about a user's activities and the interactions they have had with other Wikipedians.

<sup>13</sup> For example, edits related to Mohamed Homos, an Egyptian midfielder who usually goes by "Homos" are often mistaken for vandalism by Huggle users.

**Sorting by undesirable only.** Counter-vandalism tools like Huggle only show vandal fighters the worst parts of Wikipedia, which includes not only errors, spam, and nonsense, but also hate speech, shock images, and aggressive trolling. By showing these users only the most suspicious newcomer edits, vandal fighters get a distorted perception of the value of newcomers to the project.

Snuggle reverses this strategy by setting the default sort order to bring attention to good newcomers and their activities first. With this, Snuggle both enables and encourages its users to see the value that newcomers bring to the project. Incidentally, this sort order also encourages Snuggle’s users to critique the practices encouraged by counter-vandalism tools by juxtaposing the activities of desirable newcomers with the negative reactions they receive.

**Primarily supporting negative reactions.** Huggle users are afforded two responses to the edits that the user interface presents: pass or reject and send a warning. There’s no feature for saying “thank you” for good edits or even re-writing edits that would be good contributions if they were only formatted correctly. In chapter 4 I brought attention to the potential for these limited affordances in directing user behavior and showed a dramatic growth in the rate of both rejection of desirable newcomers’ edits and posting of warnings when Huggle was introduced.

In Snuggle the available actions that affect a newcomer are ordered purposefully. First, users are provided the means to send a personalized message – a practice which has been declining since the introduction of counter-vandalism tools[32], but has also been shown to be a motivating factor in similar circumstances[13]. Next, users are provided the means to send an invitation to the Teahouse, a space intended to be welcoming of newcomers who need help. Finally, Snuggle still does provide the means to report vandals to the administrators of Wikipedia so that they can be banned from editing if necessary.

## 6.7 INTERVIEW STUDY

I did a study that combined guided use of Snuggle and semi-structured interviews. The study served multiple goals:

- it evaluated how good Snuggle was at helping Wikipedians identify newcomers in need of help;
- it informed my thinking about the practice of socialization in Wikipedia;
- it provided an occasion for Wikipedians to reflect on their socialization practices, thus yielding an opportunity to evaluate Snuggle as a successor system.

I recruited Wikipedians who were recently active in projects aimed at positive interactions between experienced editors and newcomers (e.g., Adopt-a-user, the Teahouse and Articles for Creation<sup>14</sup>). I posted 70 invitations, received 25 responses, and con-

<sup>14</sup> a space in Wikipedia designed to help new editors create encyclopedic articles

ducted 14 semi-structured interviews (at which point I saw a convergence of themes). Three of the participants had used Snuggle before and 11 had not.

I performed the interviews using Google Hangout and Skype; I used their screen sharing features to virtually look over participants' shoulders while they used Snuggle. The semi-structured interview and guided use session occurred in three phases. First, I asked a set of questions designed to check my assumptions (e.g., "How often do newcomers run into trouble and not know where to go for help?") and extend my understanding of current mentoring practices (e.g., "Where do you interact with newcomers?"). Next, I had participants load the Snuggle interface and gave them a high level overview of the system using a standard script that described Snuggle's user dossier lists, but did not instruct the participants about how to use Snuggle. Then, I asked them to perform a task: identify a desirable newcomer in need of help using Snuggle. Finally, I concluded with a discussion of their strategy for performing the task, their opinion of Snuggle's user interface, how they felt about categorizing newcomers, and when they might consider performing each of the wiki actions.

#### 6.7.1 *Results & discussion*

In presenting these results, I include quotations both from the interviews and from conversations started by Wikipedians on the Snuggle discussion forum.

##### 6.7.1.1 *Mentors agree with my conceptualization of the problem*

The participants agreed that new editors regularly run into trouble and don't know where to go to get help. When asked how common an experience this was for newcomers:

#6: "I think it's very common. If they start doing anything, they're going to run into trouble. [...] It could be making mistakes. It could be an editor exerting ownership. They might put their talk post at the top of the page – these little rules that no one knows. Eventually you're going to do something wrong."

Most also agreed that there's no good way to find these newcomers. Some interviewees had some round-about strategies. When asked how he finds these newcomers, one of the participants explained how he used STiki, a counter-vandalism tool, to find newcomers who are making mistakes and then return to the wiki to offer them support.

##### 6.7.1.2 *Snuggle effectively supports the identification task*

Participants understood the Snuggle UI without guidance. All participants successfully used the interface to identify a newcomer they thought needed assistance within seconds of being asked to do so.

Without fail, every Snuggle user used the trace icons (see figure 34d) when looking for a newcomer in need of help. Most looked for desirable newcomers with warnings. Some looked for prolific newcomers with no messages at all. Many also de-prioritized newcomers who already had a positive social interaction (e.g., an invitation to the Teahouse).

Surfacing these traces was essential to the usefulness of Snuggle as a means to find newcomers in need of help. When discussing talk icons and looking through the list of users, interviewee #6 commented that, “Welcome is obvious. Vandal is obvious. Warning! He got a warning. It just gives me information. I know she was welcomed. I know she was invited to the Teahouse. Here’s a warning. It gives me something to work with.”

#### 6.7.1.3 *Some volunteers shy away from 1:1 interaction.*

Many users were comfortable performing actions with Snuggle. Some used Snuggle to identify newcomers and go back to Wikipedia to send messages. But others preferred not to interact with newcomers at all.

Studies of prosocial behavior in organizations found that empathy correlated strongly with citizenship behaviors directed towards specific individuals [59]. However, Finkelstein et al. observed no correlation between empathy and time spent volunteering [21]. In other words, this prosocial orientation predicts whether a volunteer will favor 1:1 interactions, but not how much time and effort they may devote.

In order to take advantage of the available resources, a good newcomer socialization system will be able to take advantage of the time and effort of both prosocial and antisocial volunteers. There are a number of such socialization tasks, e.g., manually classifying newcomers as good-faith or bad-faith and flagging good-faith newcomers in need of help.

#### 6.7.1.4 *Evidence of reflection*

**Strong reactions to undue warnings.** The interviews show that mentors were able to use the Snuggle’s user dossier to identify false positives of counter-vandalism tools and direct their support to the user. Many participants felt the need to act immediately. For example, during the task evaluation, interviewee #10 remarked, “I don’t see why this guy was reverted. [...] I don’t see how this is vandalism. This is a false positive. I’m going to go ahead and categorize him as good faith.” He then sent a message to the newcomer discussing warnings the newcomer received, offering his help and finishing his message with, “Keep up the good work.” This example demonstrates how Snuggle brings visibility to a destructive part of Wikipedians’ current socialization practices and the strong reaction that some Wikipedians have when they see an example of it.

**Complexity in categorizing.** Some participants were uncomfortable coming to a conclusion about the value of another person without substantial interaction. #7: “I will not do that [categorizing] very fast to someone. Judging people or categorizing them

before I've interacted with them or just based on a limited history is very hard." Other users were less concerned about the practice. #1: "I have no personal issues with that [categorization]. [...] You're going to form an opinion anyway." Some still saw the user dossier as a collection of actions. #12: "I suppose I felt like I was categorizing edit patterns rather than people."

**Visibility of actions.** I hoped that social translucence via the public and prominent recent activity feed (described in [Social translucence](#)) would enable the social processes necessary for use of Snuggle to develop into a robust practice. However, some participants on the Snuggle forum were worried that this was too much visibility (e.g., "I would rather not have my activity on Snuggle be too accessible."), while others welcomed it (e.g., "Given that Snuggle is a promising new application, I believe that its activities should remain completely public for the time being.").

**Who gets to use Snuggle?.** Discussion forum participants raised concerns about who would be able to use Snuggle: specifically, could newcomers do so? For example, one Wikipedian stated, "I think that whatever decision we come to, the biggest thing is to not advertise the existence of Snuggle to newbies." and another confirmed, "One issue that I see a lot lately is that we have helpers and mentors in the various help spaces [...] who are not sufficiently experienced." Some interview participants brought up similar suggestions. #9 "I think that users should be autoconfirmed at least."<sup>15</sup>

## 6.8 CONCLUSION

In this chapter, I described the design and evaluation of a novel user interface design to help solve a complex socio-technical problem in Wikipedia. I used insights from both second-wave and third-wave HCI to design and evaluate a newcomer socialization system as an effective means of critique. I'll now discuss some implications for design and implications for designers.

### 6.8.1 *Implications for design*

As a newcomer socialization system, Snuggle is designed to solve a specific social information processing problem, which does not universally exist. In this frame, Snuggle-like systems are less useful when the reactions that newcomers receive are consistent with their behavior – but who gets to decide what is and is not "consistent"? This is typically an abstract third-wave question, but I found it useful to both designing a more effective system as well as to critique ideology. While existing counter-vandalism systems were more efficient because they had one reviewer evaluating one edit, I found that a diversity of reviewers evaluating a holistic dossier revealed

---

<sup>15</sup> 'autoconfirmed' refers to a minimum experience threshold applied to newcomers who have registered their accounts at least 4 days ago have made at least 10 edits.

inconsistencies that were otherwise obscured. I used a variety of techniques and was inspired by many literatures to study and reveal these inconsistencies.

Given that these issues arose in Wikipedia alongside automated evaluation systems, systems that use algorithms for similar purposes may have similar issues and can learn from Snuggle, both as a warning for designers of new systems and as a reaction to existing systems. For example, in massive open online courses, students number in the thousands. Automation is a common response to assessment, as evaluation by hand can become as impractical as it was in Wikipedia. Recalling the case of Wikipedia's counter-vandalism tools, such automated grading tools ought to:

**SHOW THE HUMAN, NOT JUST THE ACTION.** Give a holistic view of students' activities so that graders can judge the current activity in context of their other work.

**DON'T ONLY BRING ATTENTION TO NEGATIVES.** Don't focus the grader's attention purely on incorrect answers and mistakes. Bring equal attention to the rest of their work.

**SUPPORT AND SCALE EXISTING PRACTICES.** Afford the same types of nuanced feedback currently employed by non-automated graders in well-designed traditional courses.

However, if these automated systems fail in the same way that Wikipedia's counter-vandalism systems do, a Snuggle-like tool that brings attention to inconsistencies between the desirable characteristics of students pre-assessment (e.g. time spent on material) and the assessments they receive (grades) can direct support to good students who get bad grades in the short term and enable the types of visibility necessary to change minds about what good grading practice looks like in the long term.

### 6.8.2 *Implications for designers*

In describing this work, I draw heavily from second and third-wave thinking not only to make a point, but to present my rationale for designing and evaluating Snuggle accurately. I could have written about Snuggle as either a second-wave project that benefited from third-wave thinking, or as a third-wave project that benefited from second-wave thinking. Those would not only be revisionist accounts, but would miss a valuable lesson for HCI researchers. Throughout the development of Snuggle, I focused primarily on solving a large, complicated, socio-technical problem. In my focus on figuring out how to solve this problem, I was guided by whatever way of thinking helped me move forward effectively at any given moment. I strategically deployed concepts of quantification, formalization, and information processing in ways that helped provoke critical reflection, not to make a point, but because it was my best shot at finding something that might work. I found the two ways of thinking about the design of user interfaces to be complementary – we can design powerful tools that

allow humans to work more efficiently and still be informed by and actively resist the destructive power imbalances they may create.

Part III

CONCLUSION

## 7 CONCLUSION

---

### 7.1 SUMMARY

In my work, I've taken a depth-first approach to examining a set of phenomena surrounding a specific open production system, Wikipedia. Rather than developing a particular skill set or way of understanding and applying it to many systems and contexts (breadth-first), I instead focused my studies on exploring a specific phenomena in detail and building off of past results in future analyses.

In order to explore the potential for software to support Wikipedia's open production system, I first answered several fundamental questions about how Wikipedia works and where it breaks down. My goal in this is to build the fundamental understanding necessary for generalization to be useful.

However I didn't perform this work in a vacuum. In order to make progress, I drew on results, methods and perspectives from computer science, psychology, social science, sociology, political activism and literary critique. In doing so, I used Wikipedia as a virtual petri dish for exploring the nature of mass collaboration between humans as mediated by computer systems. In situating my work within these contexts, I hope to inform my own and other's work designing and studying related systems.

#### 7.1.1 *My process*

I've found that my natural process for exploring this phenomenon can be expressed in two steps:

1. exploration: figure out what's going on (work patterns, inefficiencies, etc.) and
2. experimentation: design and evaluate software that is intended to help

Based on these two steps, I can divide the substantive chapters of this thesis into two major sections.

##### 7.1.1.1 *Step one: Exploration*

In the first three studies, I focused primarily on understanding how a crucial subsystem of Wikipedia (quality control) works and where it breaks down.

In chapter 2, I asked a fundamental question about Wikipedia's quality control system: *Under which circumstances are contributions rejected?* I found that (1) there's no evidence of a learning effect whereby editors' work becomes more acceptable over time and (2) as the psychology literature predicts, Wikipedians are plagued by a strong ownership bias that undermines their ethos of openness.

In chapter 3, I showed several consistent characteristics of Wikipedia’s quality control system as it relates to the productivity of editors who are rejected. While I was able to show evidence that editors are demotivated by rejection (especially newcomers), I also showed evidence that rejection serves as a check on the *boldness* of editors activities. Historically, rejection played a beneficial role in improving the productivity of editors. However, there was one interaction that stood out and drove my later work, newcomers were especially demotivated when reverted by highly experienced editors.

In chapter 4, I used a three-pronged approach to explore the nature of Wikipedia’s declining participation. (1) I showed evidence that Wikipedia’s quality control system has become less forgiving to newcomers despite the fact that the quality of their contributions has not decreased. (2) I showed that this less forgiving environment corresponded with the adoption of efficient quality control tools by Wikipedians in order to deal with the massive growth that threatened the viability of the project. (3) I also showed that Wikipedia’s rules of behavior (Policies & Guidelines) have calcified against the changes of newcomers. In other words, editors who joined Wikipedia after a certain point weren’t able to change the rules to address the problems that they encountered as newcomers.

#### 7.1.1.2 Step two: Experimentation

In the last two substantive chapters, I tested changes to Wikipedia’s software environment intended to improve upon the observed inefficiencies.

In chapter 5, I worked with the Wikimedia Foundation to deploy and test Article Feedback, a new, extended type of contribution medium which had interesting theoretical implications (based on social learning theory) for Wikipedia’s editor decline. I showed that by opening up this new contribution type, a massive amount of new, useful contributions would be made by readers who had not previously edited the encyclopedia. But in order to take advantage of this new contribution type, automated quality control was necessary in order to make the work of moderating the submissions worth Wikipedians’ time away from editing.

And finally, in chapter 6, I designed an intelligent user interface with the intention of rectifying Wikipedia’s most pressing problem, the declining retention of desirable new editors. I designed a novel algorithmic approach to sorting newcomers by their desirability, tied the interface into Wikipedia’s social structure and incorporated elements designed to enable the necessary social processes to build a practice around the use of this new tool. And in order to design this new interface and evaluate its effectiveness, I brought together techniques from two “waves” of HCI that are widely seen as incommensurable.

## 7.2 IMPACT

A critical question at this point is, “*Where is the impact?*” What contributions has this work made to my research community, academia in general and humanity?

I didn't simply choose Wikipedia as the focus of my research because of its success and popularity. I'm primarily interested in Wikipedia because it has become very<sup>1</sup> important. I don't mean to say that lightly. Wikipedia is read by 1/10th of the planet on a monthly basis and that figure is steadily growing<sup>2</sup>. The fact that so many people are accessing an encyclopedia on a somewhat regular basis should be cause for at least reflection if not celebration.

While my initial work in this area began with pure curiosity, once I learned of the editor decline<sup>[84]</sup> at the WikiSym conference in 2009, I saw an opportunity to use my research to have a positive impact on this important project. My work since has focused on figuring out the reason for the decline (see chapters 3 & 4) and exploring potential solutions (see chapters 5 & 6).

A few concrete examples of the impact of my work can be found in Wikipedia and the non-profit company that keeps the website online, the Wikimedia Foundation. My work exploring the reason for the decline has been featured on the Wikimedia Foundation blog (e.g., <sup>3</sup>) and has led to changes in the way they prioritize new features for the MediaWiki software. In discussing my work with others at the Foundation, Steven Walling (Product Manager) wrote<sup>4</sup>:

Aaron was a major contributor to our understanding of why we have a problem retaining new editors [...] Many of our key findings are from his research. Aaron also helped Dario and I formulate some of the first experiments on the team formerly known as E3 [Editor Engagements Experiments team].

The Wikimedia Foundation's E3 team's purpose is to positively affect the rate of newcomer retention in Wikipedia through changes to the user interface. This strategy was based on the methods used in my early experiments with the same goal <sup>[32]</sup>. Similarly, the set of metrics used by the organization to measure the quantity and quality of editors' work as well as their survival rate are based on metrics that I developed for my research<sup>5</sup>.

My work has impacted initiatives within Wikipedia's volunteer community as well. For example, the Teahouse project geared toward improving the resources available for new editors was pursued due in large part to the results presented in chapter 4 (e.g., <sup>[64]</sup>). Further, a WikiProject<sup>6</sup> for solving the editor retention problem<sup>7</sup> was created as a reaction to my research and has officially endorsed the use of the Snuggle software tool discussed in chapter 6. Individual Wikipedians have found inspiration in my work too. In searching for Wikipedians to interview about their mentoring activities, I came

<sup>1</sup> My former advisor, John Riedl, quickly made me aware that the word "very" had no place in scholarly writing. This is one place where I humbly disagree.

<sup>2</sup> <http://reportcard.wmflabs.org/>, retrieved Sept. 19th, 2013

<sup>3</sup> <http://blog.wikimedia.org/2012/03/27/analysis-of-the-quality-of-newcomers-in-wikipedia-over-time/>

<sup>4</sup> This quotation was extracted from an internal mailing list, so I regretfully, cannot provide a useful link.

<sup>5</sup> See <http://meta.wikimedia.org/wiki/Research:Metrics> for a list.

<sup>6</sup> In Wikipedia, WikiProjects represent sub organizations within the community that focus on specific issues.

<sup>7</sup> WikiProject Editor Retention: [en.wikipedia.org/wiki/Wikipedia:WER](http://en.wikipedia.org/wiki/Wikipedia:WER)

across many who linked to and discussed the open access summary I published about the work from chapter 4<sup>8</sup>.

But Wikipedia is just one example of the type of collective good that networked computers may yet afford humanity. In the wake of the success of Wikipedia, many other wiki-based information repositories were enthusiastically constructed, and yet fizzled where Wikipedia excelled [34, 77]. In my work, I've begun the process of building the understanding necessary to identify what makes Wikipedia successful – a set of conclusions that should be applicable to other open production communities in order to improve their chance of such wide success. Further, by situating my work in the context of robust results from other fields (e.g. the ownership bias[88] and theories of commons governance[71]), I've paved the way for others to make use of my results in wider contexts.

System designers in other contexts can learn from my work. For example,

- members of open production communities will need to have the means to efficiently perform quality control work (see chapter 5), but there is a seemingly inevitable tradeoff between efficient quality control and negative effects on the “controlled”(see chapter 4), so designers should be cautious to design the system so that it doesn't cause long term, social issues (see chapter 6);
- system designers should expect an ownership bias when peers are afforded the ability to affect each other's work (see chapter 2);
- newcomers are most likely to be demotivated by rejection from a quality control system, but this rejection can also be an important learning experience (see chapter 3).

All things considered, even if my work does not generalize to other systems and is only relevant to Wikipedia, it's a worthy place to have focused my impact.

### 7.3 FUTURE WORK

This dissertation represents an exploration of the health of one open production system. The studies presented are only the first steps in developing a comprehensive set of theories about constructing and maintaining healthy open production communities. Presently, there are few wiki-based open production systems that match the success of Wikipedia. It has been my goal throughout this work to explore why Wikipedia has become so successful, to identify its inefficiencies and test effectiveness of potential solutions.

A strong motivator to my work has been the lack of a clear solution to Wikipedia's declining editor population. At the time, there was little more than speculation available to tell them whether the decline was even a problem, what the might be causing it, and what solutions were available. I brought data analysis, experimentation and

<sup>8</sup> For example, <https://en.wikipedia.org/wiki/User:Adjwilley?oldid=566622137>

the scientific method to shed light on this type of problem and potential solution. As discussed in the previous section, the Wikimedia Foundation and community have benefited substantially from the insights my work brought.

But the Wikimedia Foundation and I have been lucky that we managed to build a good working relationship. But what about the next community? What if the problems take a slightly different shape and the solution that works in the Wikipedia of today doesn't work in the Wikipedia, Open Street Map, Zooniverse or other open production community of tomorrow?

This is where unified theory and approachable reference material can have a substantial impact on the viability of open production systems generally. System designers of the future should not need to partner with a researcher (arguably a set of researchers as there are many others exploring Wikipedia's practical concerns) in order to figure out why their community is behaving the way that it is and what can be done about it. Systems designers would be empowered by the availability of a resource that offers guidance in general terms applicable to many different forms of open production.

I imagine a sort of hand-book of software based open production communities that collects the theory about three aspects of these systems:

- the structure of effective open production systems,
- common problems that are likely to occur in open production systems, and
- solutions to such common problems.

While Wikipedia is an excellent case study for open production, it is only just an example of such systems. In order to gather a unifying theory that is general applicable to open production systems, the study of other systems both successful and unsuccessful is necessary. I argued in the introduction to this thesis that one of the reasons that Wikipedia is such an advantageous space for exploration is the mountain of other research categorizing the perceptions, behavior, and structure of the this particular community. Future work in this direction will have to deal with exploring open production systems that are less well documented.

Despite these difficulties, I find it inspiring to imagine a world where collective action at the scale of Wikipedia is well understood and supported by networked computer systems.

## BIBLIOGRAPHY

---

- [1] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW '07*, pages 261–270. ACM, 2007. ISBN 978-1-59593-654-7. doi: <http://doi.acm.org/10.1145/1242572.1242608>.
- [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. doi: 10.1046/j.1365-2265.1999.00760.x.
- [3] J. Antin and C. Cheshire. Readers are not free-riders: reading as a form of participation on wikipedia. In *CSCW '10*, pages 127–130. ACM, 2010. doi: <http://dx.doi.org/10.1145/1718918.1718942>.
- [4] Linda Argote. *Organizational Learning: Creating, Retaining, and Transferring Knowledge*. Kluwer Academic Publishers, Norwell, MA, USA, 1999. ISBN 0792384202.
- [5] Gerard Beenen, Kimberly Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. Using social psychology to motivate contributions to online communities. In *CSCW '04*, pages 212–221. ACM, 2004. doi: 10.1145/1031607.1031642.
- [6] Ivan Beschastnikh, Travis Kriplean, and David W McDonald. Wikipedian self-governance in action: Motivating the policy lens. In *ICWSM '08*. AAAI, 2008.
- [7] Nicolas Bettenburg, Sascha Just, Adrian Schröter, Cathrin Weiss, Rahul Premraj, and Thomas Zimmermann. What makes a good bug report? In *SIGSOFT '08*, pages 308–318. ACM, 2008.
- [8] Ulrik Brandes and Jürgen Lerner. Visual analysis of controversy in user-generated encyclopedia. *Information Visualization*, 7:34–48, 2008.
- [9] Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: Transformation of participation in a collaborative online encyclopedia. In *GROUP '05*, pages 1–10. ACM, 2005. doi: 10.1145/1099203.1099205.
- [10] B. Butler, L. Sproull, S. Kiesler, and R. E. Kraut. Community effort in online groups: Who does the work and why? *Human-Computer Interaction Institute*, page 90, 2007.
- [11] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: The nature and roles of policies and rules in Wikipedia. In *CHI '08*, pages 1101–1110. ACM, 2008. doi: 10.1145/1357054.1357227.

- [12] Michelene T. H. Chi, Robert Glaser, and Ernest Rees. Expertise in problem solving. Technical report, Pittsburgh Univ., PA. Learning Research and Development Center, 1981. URL <http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED215899>.
- [13] Boreum Choi, Kira Alexander, Robert E. Kraut, and John M. Levine. Socialization tactics in Wikipedia and their effects. In *CSCW '10*, pages 107–116. ACM, 2010. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718940.
- [14] S. Cole, J. R. Cole, and G. A. Simon. Chance and consensus in peer review. *Science*, 214(4523):881–886, 1981.
- [15] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Suggestbot: using intelligent task routing to help people find work in wikipedia. In *IUI '07*, pages 32–41. ACM, 2007.
- [16] Jill P Dimond, Michaelanne Dye, Daphne Larose, and Amy S Bruckman. Hol-laback!: The role of storytelling online in a social movement organization. In *CSCW '13*, pages 477–490. ACM, 2013.
- [17] Pierpaolo Dondio, Stephen Barrett, Stefan Weber, and Jean Seigneur. Extracting trust from domain analysis: A case study on the Wikipedia project. In Laurence Yang, Hai Jin, Jianhua Ma, and Theo Ungerer, editors, *Autonomic and Trusted Computing*, volume 4158, pages 362–373. Springer Berlin, 2006.
- [18] N. Ducheneaut. Socialization in an open source software community: A socio-technical analysis. In *CSCW '05*, pages 323–368. ACM, 2005.
- [19] Michael D. Ekstrand and John T. Riedl. rv you're dumb: Identifying discarded work in wiki article history. In *WikiSym '09*, pages 4:1–4:10. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641317.
- [20] Thomas Erickson and Wendy A. Kellogg. Social translucence: An approach to designing systems that support social processes. *ACM Trans. Comput.-Hum. Interact.*, 7(1):59–83, 2000. ISSN 1073-0516. doi: 10.1145/344949.345004.
- [21] Marcia A Finkelstein and Michael T Brannick. Applying theories of institutional helping to informal volunteering: Motives, role identity, and prosocial personality. *Social Behavior and Personality*, 35(1):101–114, 2007.
- [22] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. Values at play: Design tradeoffs in socially-oriented game design. In *CHI '05*, pages 751–760. ACM, 2005.
- [23] Heather Ford and R Stuart Geiger. Writing up rather than writing ddown: Becoming wikipedia literate. In *WikiSym '12*, page 16. ACM, 2012.
- [24] Andrea Forte and Amy Bruckman. Scaling consensus: Increasing decentralization in wikipedia governance. In *HICSS '08*, pages 157–166. IEEE, 2008.

- [25] Andrea Forte, Vanesa Larco, and Amy Bruckman. Decentralization in wikipedia governance. *Journal Management Information Systems*, 26(1):49–72, 2009.
- [26] Rich Gazan. Redesign as an act of violence: Disrupted interaction patterns and the fragmenting of a social q&a community. In *CHI '11*, pages 2847–2856. ACM, 2011. doi: 10.1145/1978942.1979365.
- [27] R. Stuart Geiger. *Critical Point of View: A Wikipedia reader*, chapter The Lives of Bots, pages 78–93. Institute of Network Cultures, Amsterdam, NL, 2011.
- [28] R. Stuart Geiger and Aaron Halfaker. When the levee breaks: Without bots, what happens to Wikipedia’s quality control processes? In *WikiSym '13*. ACM, 2013.
- [29] R. Stuart Geiger and Aaron Halfaker. Using edit sessions to measure participation in wikipedia. In *CSCW '13*. ACM, 2013.
- [30] R. Stuart Geiger and David Ribes. The work of sustaining order in Wikipedia: The banning of a vandal. In *CSCW '10*, pages 117–126. ACM, 2010.
- [31] R Stuart Geiger and David Ribes. Trace wthnography: Following coordination through documentary practices. In *HICSS '11*, pages 1–10. IEEE, 2011.
- [32] R. Stuart Geiger, Aaron Halfaker, Maryana Pinchuk, and Steven Walling. Defense mechanism or socialization tactic? In *ICWSM '12*, Palo Alto, CA, USA, 2012. AAAI.
- [33] Jim Giles. Internet encyclopedias go head to head. *Nature*, 438, dec 2005.
- [34] Jonathan Grudin and Erika Shehan Poole. Wikis at work: Success factors and challenges for sustainability of enterprise wikis. In *WikiSym '10*, pages 5:1–5:8. ACM, 2010. doi: 10.1145/1832772.1832780.
- [35] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599, 1988.
- [36] Steve Harrison, Deborah Tatar, and Phoebe Sengers. The three paradigms of hci. In *Alt. CHI '07*, pages 1–18, 2007.
- [37] Edwin Hutchins. *Cognition in the Wild*, volume 262082314. MIT press Cambridge, MA, 1995.
- [38] Lilly C Irani and M Silberman. Turkoption: Interrupting worker invisibility in amazon mechanical turk. In *CHI '13*, pages 611–620. ACM, 2013.
- [39] Amy C. Justice, Mildred K. Cho, Margaret A. Winker, Jesse A. Berlin, and Drummond Rennie. Does masking author identity improve peer review quality? *JAMA*, 280(3):240–242, July 1998.

- [40] Steven J. Karau and Kipling D. Williams. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4):681–706, 1993.
- [41] Steven J. Karau and Kipling D. Williams. *Groups at Work: Theory and research*, chapter Understanding Individual Motivation in Groups: The Collective Effort Model, pages 113–141. Lawrence Erlbaum Associates, Inc., 2001.
- [42] Katz. Luring the lurkers. <http://news.slashdot.org/story/98/12/28/1745252/luring-the-lurkers>, Dec 1998.
- [43] Brian Keegan and Darren Gergle. Egalitarians at the gate: One-sided gatekeeping practices in social media. In *CSCW '10*, pages 131–134. ACM, 2010. doi: 10.1145/1718918.1718943.
- [44] Aniket Kittur and Robert E. Kraut. Harnessing the wisdom of crowds in Wikipedia: Quality through coordination. In *CSCW '08*, pages 37–46. ACM, 2008. ISBN 978-1-60558-007-4. doi: 10.1145/1460563.1460572.
- [45] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in wikipedia. In *CHI '07*, pages 453–462. ACM, 2007. doi: 10.1145/1240624.1240698.
- [46] Andrew J. Ko and Parmit K. Chilana. How power users help and hinder open bug reporting. In *CHI '10*, pages 1665–1674. ACM, 2010.
- [47] Michel Krieger, Emily Margarete Stark, and Scott R. Klemmer. Coordinating tasks on the commons: Designing for personal goals, expertise and serendipity. In *CHI '09*, pages 1485–1494. ACM, 2009. doi: 10.1145/1518701.1518927.
- [48] Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. Community, consensus, coercion, control: Cs\*w or how policy mediates mass participation. In *GROUP '07*, pages 167–176. ACM, 2007. ISBN 978-1-59593-845-9. doi: 10.1145/1316624.1316648.
- [49] Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. Articulations of WikiWork: Uncovering valued work in wikipedia through barnstars. 2008. doi: 10.1145/1460563.1460573.
- [50] Shyong (Tony) K. Lam and John Riedl. Is wikipedia growing a longer tail? In *GROUP '09*, pages 105–114. ACM, 2009. doi: 10.1145/1531674.1531690.
- [51] Shyong (Tony) K. Lam, Jawed Karim, and John Riedl. The effects of group composition on decision quality in a social production community. In *GROUP '10*, pages 55–64. ACM, 2010. doi: 10.1145/1880071.1880083.

- [52] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. Wp:clubhouse?: An exploration of Wikipedia's gender imbalance. In *WikiSym '11*, pages 1–10. ACM, 2011. doi: 10.1145/2038558.2038560.
- [53] Cliff Lampe and Paul Resnick. Slash(dot) and burn: Distributed moderation in a large online conversation space. In *CHI '04*, pages 543–550. ACM, 2004.
- [54] Michaël R. Laurent and Tim J. Vickers. Seeking health information online: Does wikipedia matter? *Journal of the American Medical Informatics Association*, 16:471–479, 2009.
- [55] J. Lave and E. Wenger. *Situated learning: Legitimate peripheral participation*. Cambridge University Press, 1991.
- [56] Stan Lee. *Amazing Fantasy #15*. Marvel Comics, 1962.
- [57] Jay Lyman. Wikipedia co-founder palnning new expert-authored site, sep 2006. URL <http://www.crmbuyer.com/story/53137.html?wlc=1235722017&wlc=1243220132>.
- [58] PD Magnus. Early response to false claims in Wikipedia. *First Monday*, 13(9), 2008.
- [59] Bonnie L McNeely and Bruce M Meglino. The role of dispositional and situational antecedents in prosocial organizational behavior: An examination of the intended beneficiaries of prosocial behavior. *Journal of applied psychology*, 79(6): 836, 1994.
- [60] John H. Miller and Scott E. Page. *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, 2007. ISBN 9780691127026.
- [61] Audris Mockus, Roy T. Fielding, and James Herbsleb. A case study of open source software development: The Apache server. In *ICSE'00*. ACM, 2000. ISBN 1-58113-206-9. doi: 10.1145/337180.337209.
- [62] Jonathan T. Morgan and Mark Zachry. Negotiating with angry mastodons: The Wikipedia policy environment as genre ecology. In *GROUP '10*, pages 165–168. ACM, 2010. doi: 10.1145/1880071.1880098.
- [63] Jonathan T. Morgan, Robert M. Mason, and Karine Nahon. Negotiating cultural values in social media: A case study from wikipedia. In *HICSS '12*, pages 3490–3499, Washington, DC, USA, 2012. IEEE Computer Society. doi: 10.1109/HICSS.2012.443.
- [64] Jonathan T Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. Tea and sympathy: crafting positive new user experiences on wikipedia. In *CSCW '13*, pages 839–848. ACM, 2013.

- [65] Michael J Muller. *Human-computer interaction: Development process*, chapter Participatory design: The Third Space in HCI, pages 165–185. CRC Press, 2003.
- [66] David R. Musicant, Yuqing Ren, James A. Johnson, and John Riedl. Mentoring in wikipedia: a clash of cultures. In *WikiSym '11*, pages 173–182. ACM, 2011. doi: 10.1145/2038558.2038586.
- [67] Bonnie A Nardi. *Context and consciousness: Activity theory and human computer interaction*. The MIT Press, 1996.
- [68] B. Nonnecke and J. Preece. Why lurkers lurk. In *Americas Conference on Information Systems*, 2001.
- [69] B. Nonnecke and J. Preece. Silent participants: Getting to know lurkers better. *From usenet to CoWebs*, pages 110–132, 2003.
- [70] Donald A Norman. *Cognitive artifacts*. Department of Cognitive Science, University of California, San Diego, 1990.
- [71] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge Univ Pr, 1990.
- [72] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: A study of power editors on Wikipedia. In *GROUP '09*, pages 51–60. ACM, 2009. doi: 10.1145/1531674.1531682.
- [73] J. Preece and B. Shneiderman. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Trans HCI* 1, 1(1):13–32, 2009.
- [74] Reid Priedhorsky, Jilin Chen, Shyong Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP'07*, Sanibel Island, Florida, USA, 2007.
- [75] Al M. Rashid, Kimberly Ling, Regina D. Tassone, Paul Resnick, Robert Kraut, and John Riedl. Motivating participation by displaying the value of contribution. In *CHI'06*, pages 955–958. ACM, 2006. doi: 10.1145/1124772.1124915.
- [76] Eric S. Raymond. The cathedral and the bazaar, 1997. URL <http://www.catb.org/~esr/writings/cathedral-bazaar/cathedral-bazaar/index.html>.
- [77] Camille Roth, Dario Taraborelli, and Nigel Gilbert. Measuring wiki viability: An empirical assessment of the social dynamics of a large sample of wikis. In *WikiSym '08*, pages 1–5. ACM, 2008.
- [78] Jodi Schneider, Alexandre Passant, and John G. Breslin. Understanding and improving wikipedia article discussion spaces. In *SAC '11*, pages 808–813. ACM, 2011. doi: 10.1145/1982185.1982358.

- [79] Sam Silverwood-Cope. Wikipedia: Page one of google uk for 99% of searches, feb 2012. URL <http://www.intelligentpositioning.com/blog/2012/02/wikipedia-page-one-of-google-uk-for-99-of-searches/>.
- [80] Adam Smith. *The Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.
- [81] Anselm L Strauss. *Negotiations: Varieties, contexts, processes, and social order*. Jossey-Bass San Francisco, 1978.
- [82] Besiki Stvilia, Michael B. Twidale, and Linda C. Smith. Information quality: Discussions in Wikipedia. *American Society for Information Science and Technology*, 59(6):983–1001, 2005.
- [83] Bongwon Suh, E.H. Chi, B.A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. In *VAST '07*, pages 163–170, Nov. 2007. doi: 10.1109/VAST.2007.4389010.
- [84] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. The singularity is not near: Slowing growth of Wikipedia. In *WikiSym '09*, pages 1–10. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641322.
- [85] Dario Taraborelli and Camille Roth. Viable web communities: Two case studies. In Guillaume Deffuant and Nigel Gilbert, editors, *Viability and Resilience of Complex Systems*, pages 75–105. Springer, 2011. doi: 10.1007/978-3-642-20423-4\_4.
- [86] Gillespie Tartelton. The politics of platforms. *New Media & Society*, 12(3):347–364, 2010.
- [87] Jennifer Thom-Santelli, Dan R. Cosley, and Geri Gay. What’s mine is mine: Territoriality in collaborative authoring. In *CHI'09*. ACM, 2009. ISBN 978-1-60558-246-7.
- [88] Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, 106(4):1039–1061, 1991. doi: 10.2307/2937956.
- [89] Susan van Rooyen, Fiona Godlee, Stephen Evans, Richard Smith, and Nick Black. Effect of blinding and unmasking on the quality of peer review: A randomized trial. *JAMA*, 280(3):234–237, July 1998. doi: 10.1001/jama.280.3.234.
- [90] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI'04*. ACM, 2004. ISBN 1-58113-702-8. doi: 10.1145/985692.985765.
- [91] Fernanda B. Viégas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. Talk before you type: Coordination in Wikipedia. In *HICSS '07*, page 78, Washington, DC, USA, 2007. IEEE Computer Society. ISBN 0-7695-2755-8. doi: 10.1109/HICSS.2007.511.

- [92] G. Von Krogh, S. Spaeth, and K.R. Lakhani. Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, 32(7): 1217–1241, 2003.
- [93] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan Lauw. On ranking controversies in Wikipedia: Models and evaluation. In *WSDM'08*. ACM, 2008. ISBN 978-1-59593-927-9. doi: 10.1145/1341531.1341556.
- [94] Jimmy Wales. Wikipedia sociographics. 21st Chaos Communication Congress, December 2004. URL <http://ccc.de/congress/2004/fahrplan/event/59.en.html>.
- [95] Rick Wash and Cliff Lampe. The power of the ask in social media. In *CSCW '12*, pages 1187–1190. ACM, 2012.
- [96] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *iConference '11*, pages 122–129. ACM, 2011.
- [97] Andrew G West, Sampath Kannan, and Insup Lee. Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata? In *EuroSys '10*, pages 22–28. ACM, 2010.
- [98] Andrew G. West, Avantika Agrawal, Phillip Baker, Brittney Exline, and Insup Lee. Autonomous link spam detection in purely collaborative environments. In *WikiSym '11*, pages 91–100. ACM, 2011. doi: 10.1145/2038558.2038574.
- [99] Wikimedia. Editor trends study. [http://strategy.wikimedia.org/wiki/Editor\\_Trends\\_Study](http://strategy.wikimedia.org/wiki/Editor_Trends_Study), March 2011.
- [100] D.M. Wilkinson. Strong regularities in online peer production. In *Ecommerce '08*, pages 302–309. ACM, 2008.
- [101] Honglei Zeng, Maher A. Alhossaini, Richard Fikes, and Deborah L. McGuinness. Mining revision history to assess trustworthiness of article fragments. In *CollaborateCom '06*, pages 1–10, nov. 2006. doi: 10.1109/COLCOM.2006.361890.
- [102] X.M. Zhang and F. Zhu. Intrinsic motivation of open content contributors: The case of Wikipedia. *Workshop on Information Systems and Economics*, 2006.

Part IV

APPENDIX

## DECLARATION

---

I, Aaron Halfaker, hereby certify that this thesis, has been written by me, that it is the record of work carried out by me, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September, 2006 and as a candidate for the degree of PhD in Computer Science in December, 2012; the higher study for which this is a record was carried out in the University of Minnesota between 2007 and 2013.

I received assistance in the writing of this thesis in respect of language, grammar, spelling and syntax, which was provided by my advisor, Loren Terveen.

*Minneapolis, MN, December, 2013*

---

Aaron Halfaker