The Rise and Decline of an Open Collaboration System: How Wikipedia's reaction to popularity is causing its decline

Aaron Halfaker[1]	R. Stuart Geiger[2]	Jonathan Morgan[3]	John Riedl[1]
halfaker@cs.umn.edu	rsg@berkely.edu	jmo25@uw.edu	riedl@umn.edu

1. GroupLens Research, University of Minnesota
2. School of Information, University of California, Berkeley
3. Human Centered Design & Engineering, University of Washington

Abstract

Open collaboration systems like Wikipedia need to maintain a pool of volunteer contributors in order to remain relevant. Wikipedia was created through a tremendous number of contributions by millions of contributors. However, recent research has shown that the number of active contributors in Wikipedia has been declining steadily for years, and suggests that a sharp decline in the retention of newcomers is the cause. This paper presents data that show that several changes the Wikipedia community made to manage quality and consistency in the face of a massive growth in participation have ironically crippled the very growth they were designed to manage. Specifically, the restrictiveness of the encyclopedia's primary quality control mechanism and the algorithmic tools used to reject contributions are implicated as key causes of decreased newcomer retention. Further, the community's formal mechanisms for norm articulation are shown to have calcified against changes – especially changes proposed by newer editors.

Introduction and related work

Open collaboration systems like Wikipedia require a large pool of volunteer contributors. Without volunteers to occupy necessary roles, these systems would cease to function. Like any volunteer community, open collaboration 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 (Wilkinson, 2008; Panciera, 2009).

The success of an open collaboration project appears to be highly correlated with the number of participants it maintains. Projects that fail to recruit and retain new contributors tend to die quickly (Ducheneaut, 2005). In order to maintain a pool of contributors, newcomers must be continually socialized into the organization. Some newcomers must move from the periphery of the community to the center (Bryant, 2005).

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 56,400 active editors. The work of this massive group has propelled the encyclopedia to a high level of quality and completeness (Giles, 2005). Suh et al. (2009) describe this growth as a self-reinforcing mechanism: as Wikipedia became more valuable, the project attracted more contributors to increase its value.

Then, at the beginning of 2007, things changed. Participation entered a period of

decline[1]. Why? Recent research suggests different explanations. Suh et al. (2009) argue that the decline could be the result of increasing completion of articles in the context of a population model. 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[2].

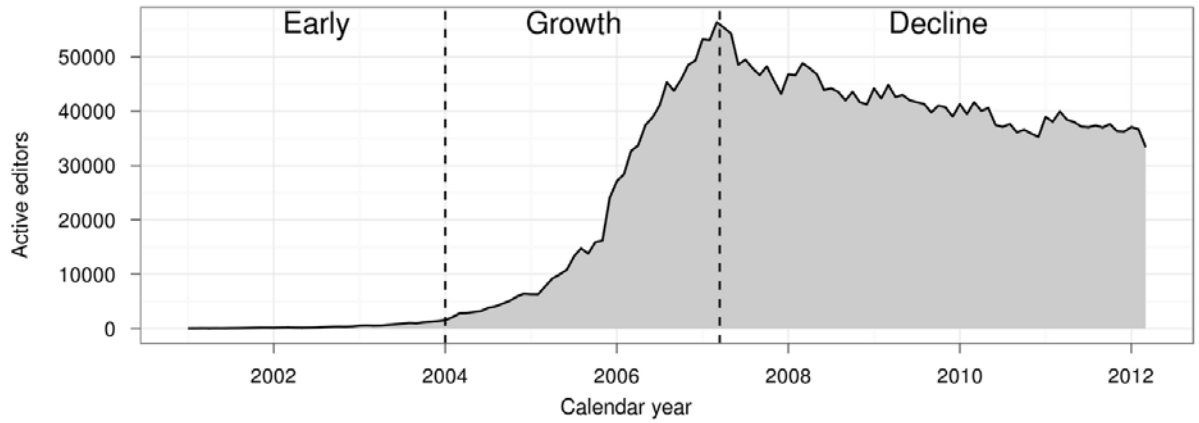

Figure 1. The English Wikipedia's editor decline. The number of active, registered editors (>= 5 edits/month) is plotted over time.

Other researchers point to failed socialization systems. Indeed, evidence suggests that it is difficult for newcomers to find work to do (Krieger, 2009) and to discover where to ask for help. 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 (Choi et al., 2010). Wikipedians have organized mentoring systems to support socialization, but they fail to serve most newcomers (Musicant, 2011).

Also, the editing community could simply be "right-sizing". Perhaps now that the main 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, underrepresented 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 (Lam, 2011). Until editors are representative of the population of potential contributors, it is difficult to argue that the socialization practices are sufficiently effective.

In this paper we define a type of Wikipedia editor that we call a *desirable newcomer*. The first few edits of these newcomers indicate that they are trying to contribute productively (i.e. acting in good-faith) and, therefore, likely will become valuable contributors if they remain in the community. We show empirically that, while the proportion of desirable newcomers who arrive at Wikipedia has been holding steady in recent years, a decreasing fraction of these newcomers survive past their initial contributions. We demonstrate that the decline has been caused, at least in part, by the Wikipedia community's reactions to the enormous influx of contributors between 2004 and 2007. In order to maintain quality and efficiency during this period, the community's views toward the goals of the project changed. These new views were instantiated

[1] http://strategy.wikimedia.org/wiki/Editor_Trends_Study
[2] http://enwp.org/WP:FAS - http://enwp.org/WP:GAS

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 rudely greeted 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, we implicate Wikipedia's primary quality control mechanism (Stvilia, 2005), 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, we show how algorithmic tools, which 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, we show how calcification has made Wikipedia's policy environment less adaptable and increased the difficulty of contributing to community rules – especially for newcomers.

Motivation & Hypotheses

Rejection of newcomers

Stvilia et al. (2005) 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. This review system is apparently effective at producing value.

Yet Halfaker et al. (2011) found 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. (2009), 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 quality of the rejecting edit 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. (2009) 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. In a related study, Halfaker et al. showed that editors were more likely to get into conflict when editing the same parts of articles (Halfaker, 2009).

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 certainly changed over time. While presenting at Wikimania in 2006, Jimmy Wales urged Wikipedians to change their focus from quantity to quality. This 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 (Lam, 2009).

There are also external pressures for Wikipedia to tighten its review process. After high profile cases of libel (e.g. the Seigenthaler libel incident[3]), 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. In an analysis performed by Halfaker[4], recent newcomers were shown to be 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.

We suspect that the increased rates of rejection are explained by changes in the way that Wikipedia deals with damage 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, we 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.

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 repetitive activities

[3] http://en.wikipedia.org/wiki/Wikipedia_biography_controversy
[4] http://meta.wikimedia.org/wiki/Research:Newbie_reverts_and_article_length

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. After years of iteration, vandal fighter bots are in wide use in mid-2012. *ClueBot NG* 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 an editor's most recent contributions to an article and send them a pre-written message telling them 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"[5] 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 (Kittur, 2007; Priedhorsky, 2007). 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[6]. A related study by Geiger et al. (2012) showed that these algorithmic tools have been taking an increasing role in "welcoming" newcomers via warning messages. 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, we 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 editing patterns they encourage.

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, we 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.

[5] "diff" refers to the visual presentation of the changes made by a single edit to an article.

[6] http://meta.wikimedia.org/wiki/Research:The_Speed_of_Speedy_Deletions

- The use of algorithmic tools to revert newcomer edits significantly increases the negative effect of rejection on desirable newcomer retention.

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 (Butler 2008), and rule creation and enforcement became increasingly decentralized (Beschastnikh 2008, Forte 2009).

The trends towards decentralization and norm formalization in Wikipedia governance may have been natural and healthy responses to community growth (Forte 2009). Formally documenting community practices facilitated wider dissemination in the expanding community, and new rules were created to meet emergent 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 (Morgan 2010), and may be widely known and highly cited (such as the *bold, revert, discuss* essay referred to below).

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 (Viegas 2007). Decentralized policy creation and enforcement allowed policies to reflect current community concerns as more editors – and, increasingly, newer editors – began to write and cite policies (Beschastnikh 2008). These findings have led researchers (Viegas 2007, Forte 2009) to characterize growth-era Wikipedia as an example of successful commons-based governance (Ostrom 1990) because policies *reflect local circumstances, are flexible enough to change in response to emergent 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 both decentralization and norm formalization have slowed. For example, decentralization has its limits: senior editors tend to have greater 'power of interpretation' over policy (Kriplean 2007, Morgan 2012) and greater control of community processes (Keegan 2010) than newer editors. And the institution of an official peer review process for new policy proposals in 2005 may have slowed new policy creation (Forte 2009). Furthermore, more recent analysis[7] 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 policy than ever before.

Although policies 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 technologically-embedded processes that are no longer open to re-negotiation. If decentralization in governance and dynamic norm formalization were key to Wikipedia's successful socialization of new members during the growth period, we suspect that policy *calcification* and increasing *centralization* of policy interpretation may negatively affect the retention rate of desirable newcomers.

[7] http://meta.wikimedia.org/wiki/Research:WikiPride

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, we 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 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 non-binding, informal norms (essays), where their contributions are significantly less likely to be rejected.

Methods

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 by Panciera et al. (2009). 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. An hour is a common session timeout used in online systems to make up for the stateless nature of HTTP. We base several metrics of editor characteristics described in this section on the contributions editors' make during their first edit sessions.

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.

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?

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[8].

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)

Tracking algorithmic tools. In order to track the use of algorithmic tools, we employ various techniques described in Geiger et al. (2012). 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.

[8] 5 raters = 2 researchers + 3 Wikipedians

Conflict discussion reciprocation. In Wikipedia, one of the most longstanding and widely-cited essays is the *Bold, Revert, Discuss cycle*[9] (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, and then
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 (Viegas, 2007; Schneider, 2010). 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.

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 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[10] 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.

[9] http://en.wikipedia.org/wiki/Wikipedia:BOLD,_revert,_discuss_cycle

[10] http://dumps.wikimedia.org/other/diffdb

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

Hypothesis: Rejection & retention

Results

To explore the validity of *Hypothesis: 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 1 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 Halfaker et al. (2011) that reverts of 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 of newcomers.

All newcomers (n = 100k, AIC: 46013)				**Desirable newcomers** (n=1708, AIC: 1720)			
	Est.	**StdErr**	**Pr(>|z|)**		**Est.**	**StdErr**	**Pr(>|z|)**
(Intercept)	-1.98	0.017	< 0.001	**(Intercept)**	-1.30	0.089	< 0.001
year	-0.40	0.012	< 0.001	**year**	-0.59	0.069	< 0.001
session edits	0.18	0.009	< 0.001	**session edits**	0.24	0.064	< 0.001
deleted	-1.45	0.037	< 0.001	**deleted**	-0.80	0.217	< 0.001
reverted	-0.68	0.035	< 0.001	**reverted**	-0.50	0.173	0.004
messaged	0.54	0.027	< 0.001	**messaged**	0.68	0.127	< 0.001
tool revert	-0.67	0.062	< 0.001	**tool revert**	-2.16	1.086	0.047

Table 1. 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.

However, these results alone do not represent a good test of *Hypothesis: 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.

Figure 2 shows that, while the combined proportion of newcomers falling into the two good 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 1 year prior to the peak and decline in Wikipedia's active contributors that began in 2007 (see figure 1).

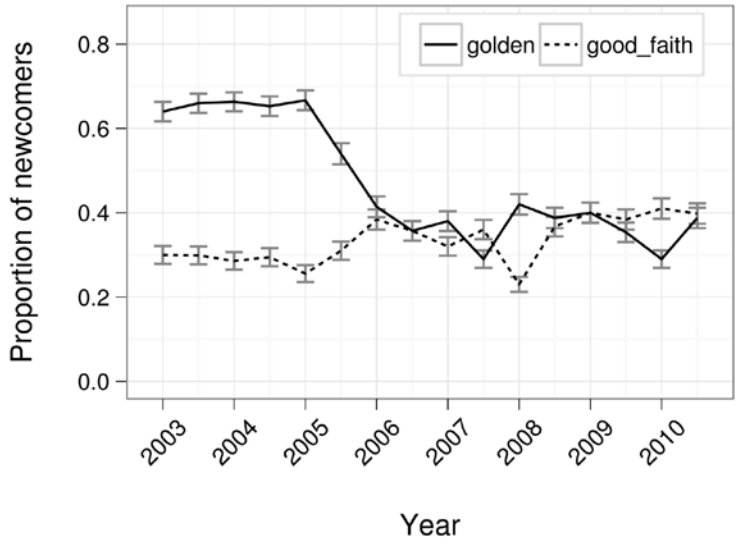

Figure 2. Quality of newcomers over time. The proportion of editors falling into the two good-faith quality categories is plotted over time.

Figure 3 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% desirable newcomers experiencing rejection in the form of a revert.

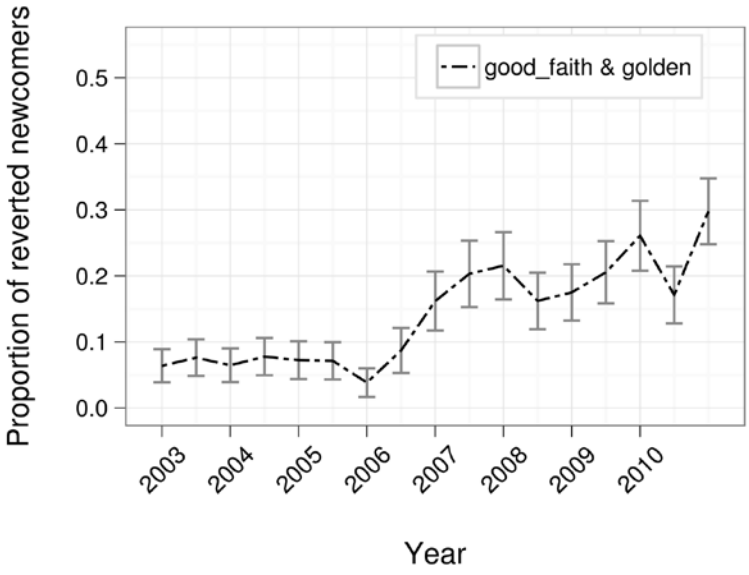

Figure 3. Reverts of desirable newcomer contributions over time. The proportion of good ("good-faith" & "golden" combined) newcomers with at least one reverted first session edit is plotted over time.

Figure 4 shows that the increasing rate of reverted desirable newcomers corresponds closely with a decline in the survival rate for desirable newcomers. 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.

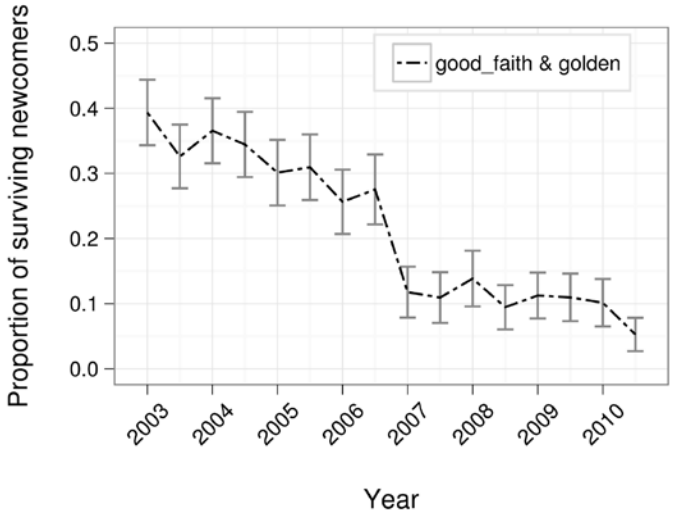

Figure 4. Survival of desirable newcomers over time. The proportion of surviving good ("good-faith" & "golden" combined) newcomers is plotted over time.

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 of newcomer retention.

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 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.

Hypothesis: Tool use & consequences

Results

Newcomer rejection. To explore the potential role of algorithmic tools as gatekeepers to the community, we built on the work of Geiger et al. (2012) by examining the rate of interaction around rejection between newcomers and the actions of algorithmic tools. Figure 5 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%.

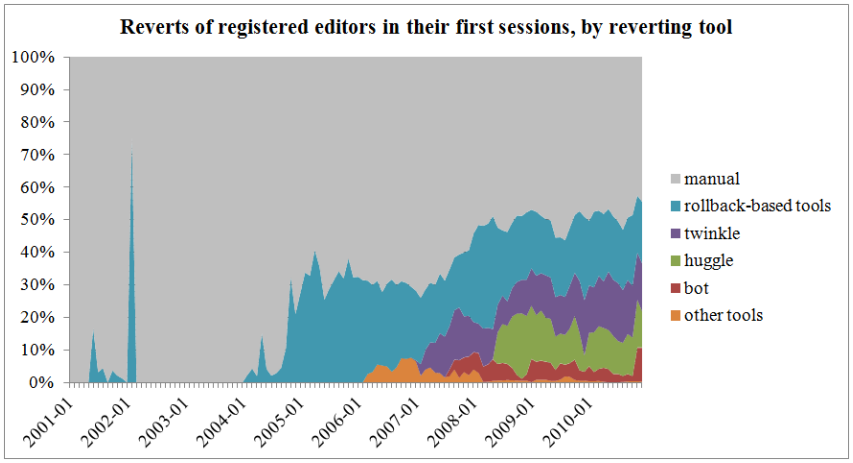

Figure 5. Use of algorithmic tools to reject newcomers edits. The proportion of rejected first session contributions is plotted for newcomers by the mechanism used for rejection over time.

The majority of tool-based rejection of newcomers came from human-computation tools – tools that borrowed human judgment. This seems reasonable given that, as reported by Geiger (2011), 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.

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.

Curiously, Figure 6 suggests that a large number of newcomers (2,250 BRD initializations from 918 unique registered editors) are attempting to enter into dialog with an algorithmic editor after being reverted by them. This might indicates a potential issue with using fully-automated bots to revert contributions.

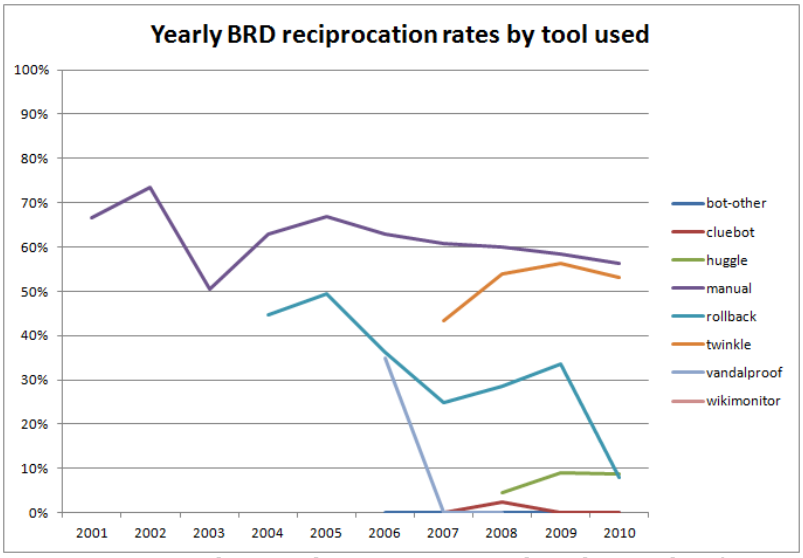

Figure 6. BRD reciprocation rates over time by tool. The proportion of newcomer BRD initiations that resulted in oldtimer reciprocation is plotted over time by the algorithmic tool used.

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 1. Both columns report a significant negative effect for *tool revert* 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.

Since the exponential growth of Wikipedia, the rate at which desirable newcomers are reverted using tools also appears to be rising. Figure 7 shows the rise of tool based rejection of newcomer contributions since starting at 0% in 2006 to 40% in 2010.

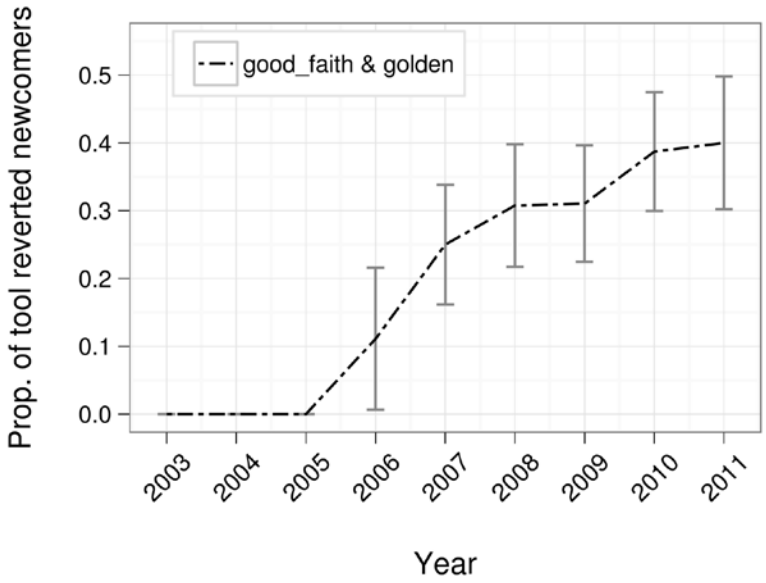

Figure 7. Rate of tool based reverts of desirable newcomers. The proportion of reverted desirable newcomers ("good-faith" & "golden" combined) who were reverted using algorithmic tools is plotted over time.

Discussion

Our analysis shows that algorithmic tools have had an increasing role in rejecting the contributions of newcomers. Given that Geiger et al. (2012) shows that these tools are also taking over the task of "welcoming" newcomers via warning messages posted on their talk page, this suggests that newcomers are increasingly rejected by and warned by 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. Together, we see this as a shift from human, personal interaction to mechanical, impersonal interaction that took place during the exponential growth of the community.

The regression analysis over *survival* shows a significant, exacerbating effect for the newcomers whose contributions were rejected using tools. The BRD 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, we suspect that this impersonal, non-communicative nature of interaction has other, possibly more difficult to measure, implications that are exacerbating the effect of rejection on retention.

Bruno Latour (1988) famously analyzed the social roles of walls, doors, and pneumatic door-closers to demonstrate the functional equivalence between humans and objects in producing social order. Considering that these algorithmic tools and agents are predominantly deployed to protect the encyclopedia from the potentially-damaging contributions of less experienced editors, it may be more appropriate to refer to such algorithms as gates instead of gatekeepers. As Latour illustrates, when tasks are delegated from humans to technologies (or vice versa), there are often dramatic shifts in social practices and responsibilities. Given how certain patterns of exclusion are embedded into Wikipedia's technological and social structure (Geiger, 2011), this highly-automated approach to policy enforcement is likely to have even farther-reaching effects on the community than those we describe in this paper.

Hypothesis: Norm formalization & calcification

Results

To explore *Hypothesis: 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 8 shows that growth of policies and guidelines began to slow in 2006, just as Forte (2009) reports. 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 (see figure 1). After that point, contribution to existing policies and guidelines begins to decline.

To look for effects of policy calcification on overall norm formalization, we compared the rate of creation and contribution to formal norms (policies and guidelines) and informal norms (essays). We find an increase in essay creation that corresponds to the decline in policy 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 8 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.

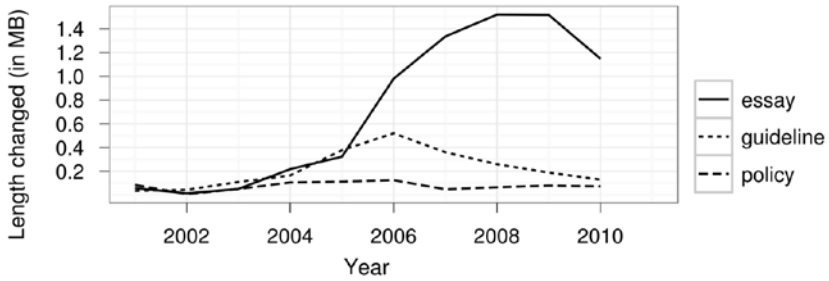

Figure 8. Norm page growth over time. The change to overall length of the three norm types is plotted by year.

To look for evidence of calcification of policies against contributions, we performed a logistic regression (described in *Methods*) to predict the rejection of new contributions to all three types of formalized norm. Table 2 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.

(n = 120535, AIC = 16801)

	Est.	StdErr	Pr(>|z|)
(Intercept)	-1.50	0.043	< 0.001
editor tenure	-0.29	0.006	< 0.001
year	0.12	0.006	< 0.001
essay	-0.38	0.135	0.005
editor tenure:essay	0.06	0.019	0.002
year:essay	-0.10	0.019	< 0.001

Table 2. The coefficients of a logistic regression over the contributions of registered editors to norm pages predicting *success* (i.e. not reverted) are presented.

Discussion

Our analysis shows that the documentation of new formal norms has declined, and it has become more difficult over time for Wikipedia editors to contribute to existing policy – especially editors from more recent cohorts. We offer the rising rate of rejection as evidence of calcification and explain the slowing growth of formal norms as the likely outcome of such a process.

We see at least two consequences of policy calcification that bear directly on newcomer socialization and retention. First, the calcification of policy is disproportionately felt by newer

editors, who see their policy edits rejected at a higher rate. This suggests that under Wikipedia's current policy 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, and increasing the power imbalance between newer and older editors. 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. As the BRD analysis above suggests, the informal norms documented in essays are trumped by formal norms embedded in bots and human computation tools.

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 (Stvilia, 2005) and it works (Giles, 2005). However, as the scale 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 formatting 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 (Geiger, 2010). Without algorithmic tools, substantially more volunteer effort would be needed to protect the encyclopedia from damage, and quality 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 successfully 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 contribute to, especially for newer 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.

While there are many lessons to be learned from the story of Wikipedia's rise and decline, we conceptualize this as a case of socio-technical gate-keeping and its consequences. 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 collaboration projects. Online communities generally must deal with how to enforce norms and regulate behavior. A variety of strategies can be taken to this effect. For example, Lampe and Resnick (2004) studied the highly 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. Another study by Gillespie (2010) examined the copyright infringement detection algorithms used by YouTube to automate the process of identifying and removing infringing context. 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 social activities that those tools afford and restrict.

Acknowledgments

Thanks to Oliver Keyes, Maryana Pinchuk and Steven Walling for their work in assessing newcomer quality and the support of the National Science Foundation, under grants IIS 09-68483 and IIS 11-11201.

Bibliography

Beschastnikh, I., Kriplean, T., McDonald, D. W. (2008). Wikipedia Self-Governance in Action: Motivating the Policy Lens. *ICWSM*.

Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia. *GROUP* (pp. 1-10)

Butler, B., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. *CHI* (pp. 1101-1110)

Choi, B., Alexander, K., Kraut, R. E., & Levine, J. M. (2010). Socialization tactics in wikipedia and their effects. *CSCW* (pp. 107-116).

Ducheneaut, N. 2005. Socialization in an Open Source Software Community: A Socio-Technical Analysis. *CSCW* (pp. 323-368).

Forte, A., Larco, V., & Bruckman, A. (2009). Decentralization in Wikipedia Governance. *Journal of MIS* 26(1), 49-72.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature*, 438(7070), 900-901.

Geiger, R. S. (2011) The Lives of Bots. In *Critical Point of View: A Wikipedia Reader*, G. Lovink and N. Tkacz, eds., Institute of Network Cultures, (pp. 78-79).

Geiger, R. S., Halfaker, A., Pinchuk, M., & Walling, S. (2012). Defense Mechanism or Socialization Tactic? Improving Wikipedia's Notifications to Rejected Contributors. *ICWSM*.

Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in Wikipedia: the banning of a vandal. *CSCW* (pp. 117-126).

Gillespie, T. (2010) The Politics of Platforms. *New Media and Society*, 12(3), 347-364.

Halfaker, A., Kittur, A., Kraut, R. E., & Riedl, J. A jury of your peers: quality, experience and ownership in Wikipedia. *WikiSym* (pp. 15:1-10).

Halfaker, A., Kittur, A., & Riedl, J. (2011). Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. *WikiSym* (pp. 163-172).

Keegan, B., Gergle, D. (2010). Egalitarians at the gate: One-sided gatekeeping practices in social media. *CSCW* (pp. 131-134).

Kittur, A., Suh, B., Pendleton, B. A., & Chi, E. H. He says, she says: conflict and coordination in Wikipedia, *CHI* (pp. 453-462)

Krieger, M., Stark, E. M., & Klemmer, S. R. (2009). Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. *CHI* (pp. 1485-1494).

Kriplean, T., Beschastnikh, I., McDonald, D. W., & Golder, S. A., (2007) Community, consensus, coercion, control: cs*w or how policy mediates mass participation. *GROUP* (pp. 167-177).
Lam, S. K., & Riedl, J. (2009). Is Wikipedia growing a longer tail?. *GROUP* (pp. 105-114).
Lam, S. K., Uduwage, A., Dong, Z., Sen, S., Musicant, D. R., Terveen, L., Riedl, J. (2011). WP:clubhouse?: an exploration of Wikipedia's gender imbalance. *WikiSym* (pp. 1-10).
Lampe, C., & Resnick, P. (2004). Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. *CHI* (pp. 543-550).
Latour, B. (1988). Mixing humans and nonhumans together: The sociology of a door-closer, *Social Problems*, 298-310.
Morgan, J. T., Mason, R. M., & Nahon, K. (2012). Negotiating Cultural Values in Social Media: A Case Study from Wikipedia. *HICSS* (pp. 3490-3499).
Morgan, J. T., & Zachry, M. (2010). Negotiating with angry mastodons: the Wikipedia policy environment as genre ecology. *GROUP* (pp. 165-168).
Musicant, D. R., Ren, Y., Johnson, J. A., & Riedl, J. (2011). Mentoring in Wikipedia: a clash of cultures. *WikiSym* (pp. 173-182).
Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action.* New York: Cambridge University Press.
Panciera, K., Halfaker, A., & Terveen, L. (2009) Wikipedians are born, not made: a study of power editors on Wikipedia. *GROUP* (pp. 51-60).
Priedhorsky, R., Chen, J., Lam, S. K., Panciera, K., Terveen, L., & Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia, *GROUP* (pp. 259-268).
Stvilia, B., Twidale, M. B., Smith, L. C., & Gasser L (2005). Information quality work organization in Wikipedia. *American Society for Information Science and Technology*, 59(6), 983-1001.
Suh, B., Convertino, G., Chi, E. H., & Pirolli, P. (2009) The singularity is not near: slowing growth of Wikipedia. WikiSym, (pp. 8:1-10).
Schneider, J., Passant, A., & Breslin, J.G. (2011). Understanding and improving Wikipedia article discussion spaces. *SAC* (pp. 808-813).
Viegas, F. B., Wattenberg, M., & McKeon, M. M. (2007). The Hidden Order of Wikipedia. OCSC (pp. 445-454).
Viegas, F. B., Wattenberg, M., Kriss, J., & van Ham, F. (2007) Talk Before You Type: Coordination in Wikipedia, *HICSS* (pp. 78-88).
Wilkinson, D. M. (2008). Strong regularities in online peer production. *ECom* (pp. 302-309).