When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes?

R. Stuart Geiger
School of Information
University of California, Berkeley
102 South Hall
Berkeley, CA

stuart@stuartgeiger.com

Aaron Halfaker
GroupLens Research
University of Minnesota
200 Union St. SE. 4-192A
Minneapolis, Minnesota

halfaker@cs.umn.edu

ABSTRACT

In the first half of 2011, ClueBot NG – one of the most prolific counter-vandalism bots in the English-language Wikipedia – went down for four distinct periods, each period of downtime lasting from days to weeks. In this paper, we use these periods of breakdown as naturalistic experiments to study Wikipedia's heterogeneous quality control network, which we analyze as a multi-tiered system in which distinct classes of reviewers use various reviewing technologies to patrol for different kinds of damage at staggered time periods. Our analysis showed that the overall time-to-revert edits was almost doubled when this software agent was down. Yet while a significantly fewer proportion of edits made during the bot's downtime were reverted, we found that those edits were later eventually reverted. This suggests that other agents in Wikipedia took over this quality control work, but performed it at a far slower rate.

Categories and Subject Descriptors

H.5.3 [**Information Systems**]: Group and Organization Interfaces—*computer-supported collaborative work*

Keywords

Wikipedia, peer production, information quality, automation, bots, software agents, socio-technical systems

1. INTRODUCTION

This paper is about one of the most active contributors to the English-language Wikipedia. Like many Wikipedians, this editor is known by a pseudonym, ClueBot NG – although others often drop the suffix for brevity. While ClueBot NG is a relative newcomer to the project, having first edited the English-language Wikipedia in November of 2010, this editor quickly gained notoriety and respect patrolling for spam and vandalism. To say that this editor is dedicated is an understatement: ClueBot NG makes thousands of edits every day, averaging around 5,000 on weekdays and about 2,500 on weekends. With over 2 million edits as of March 2013, ClueBot NG is the 6th most prolific editor to the English-language version of Wikipedia in terms of the raw number of edits made. ClueBot NG has also received dozens of barnstars [6] as tokens of appreciation from Wikipedians who are grateful for keeping their articles free of damaging content, from spam and vandalism to patent nonsense and page blanking.

However, as this user's name suggests, ClueBot NG is not entirely a human, but is instead what is referred to as a "bot" – an autonomous software program which is developed and operated by volunteers. ClueBot NG's sophisticated damage-detection algorithms are fast enough to scan every edit made to Wikipedia in real time, reverting any edit which it deems to be harmful to the encyclopedia. Built on Bayesian neural networks and trained with data about what kind of edits Wikipedians regularly revert as vandalism, the bot is designed to embody and enforce the standards and practices that constitute Wikipedian understandings of encyclopedicness. It provides a critical gatekeeping function in a peer-production community like Wikipedia, where nearly anyone with Internet access has the technical ability to edit nearly any page in any way they see fit. Without ClueBot NG working to remove undesirable content twenty-four hours a day, seven days a week, Wikipedia as both an encyclopedic text and a social organization could look quite different.

Yet like most human editors, bots take breaks from editing Wikipedia every now and then, although for somewhat different reasons. Sometimes there is a bug in the code that causes it to malfunction, and the bot will be shut down until its operator can figure out how to fix it. As bots must be scaffolded onto existing platforms, sometimes the site's APIs or data formats change in ways that are incompatible with how a bot has been designed. As bots are not built into the MediaWiki platform and instead must be run on remote machines, sometimes the computers hosting these bots break and must be fixed. A bot can go down because a developer is hosting it on their personal computer and has to move across the country A bot that is run on servers

WikiSym 2013, Aug 05–07, 2013, Hong Kong, China.

owned by a developer's school or work can go down when the servers' administrators suddenly withdraw permission. In all, there are a number of reasons why bots can suddenly become inactive, leaving their fellow human and bot editors without a valuable, yet often taken-for-granted member of their community.

In the spring of 2011, the somewhat unthinkable happened: ClueBot NG began going offline, often for days at a time. The first two outages were relatively brief: the first lasted from February 15th-18th, the second from March 13th-17th. Then, a couple weeks later, on March 29th, ClueBot NG went down again, and wasn't back until over a week later, on April 7th. Finally, after a full week of continuous counter-vandalism activity, ClueBot NG went down again on April 15th, not returning to the encyclopedia project until May 1st. Figure 1 shows the number of edits ClueBot NG made per day during the first six months of 2011, and the downtimes are clearly visible.

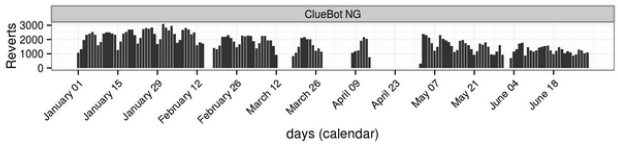

Figure 1: Edits by ClueBot NG from 1 Jan to 30 Jun, 2011

2. WIKIPEDIA'S IMMUNE SYSTEM

2.1 Previous research

Many academic and popular commentaries on Wikipedia have expressed amazement at the project's ability to revert low-quality contributions within minutes, often assuming a reviewing model in which a staggering number of Wikipedians are constantly reading entire encyclopedia articles, correcting errors as they find them. A number of Wikipedia researchers who study information quality, such as Stvilia et al.[9], have shown the many different kinds of information quality actors and practices at work, from individual editors to talk page negotiations to software agents. As Geiger and Ribes detail in their account of the banning of a vandal [2], one sector of Wikipedia's reviewing corps – "vandal fighters" – enforce quality control through complex process of distributed cognition, in which human and algorithmic agents work together to patrol the encyclopedia in near real time.

According to Geiger and Ribes, three different types of users – unassisted humans, tool-assisted cyborgs, and fully-automated bots, a typology we take from [4] – all contribute to the same overall goal of ensuring that the encyclopedia remains as free from vandalism, spam, and error as possible. What is interesting for this study is the temporal distribution of such activity: fully automated bots are able to revert the most blatantly damaging edits within seconds, often so fast enough that tool-assisted vandal fighters celebrate when they are able to beat ClueBot NG. With tools like Huggle and STiki [11], assisted humans (or cyborgs) make up a second line of defense: they are given a filtered or unfiltered set of edits to review in a live queue, and with a single click, they can instantly revert the edit in question and advance to the next one. Finally, a third line of defense is made up by editors who are interacting with Wikipedia in their web browsers, often engaging in a more traditional mode of editorial review. Some user interface extensions and in-browser functions have been developed (like Twinkle, rollback, and undo) so that editors can revert damaging edits with a single click, avoiding the tedious process of having to click the "edit" button and manually remove the offending text.

Geiger and Ribes's account is based on ethnographic fieldwork with vandal fighters, administrators, and other quality control agents in Wikipedia, and they argue that such algorithmic quality control tools are part of what makes it possible for Wikipedia to exist as an open encyclopedia that anyone can edit. ClueBot NG's downtime gives us a unique opportunity to put these claims to the test, drawing on a long tradition in the study of sociotechnical systems that uses moments of breakdown to study the role of infrastructure. [1, 8]

2.2 Operationalizing quality control activity

Before we investigated the case of ClueBot NG's downtime and its effects, we sought to develop metrics which would let us operationalize activity and outcomes in this system. We first had to develop metrics that would either confirm or complicate Geiger and Ribes's depiction of Wikipedia's vandal fighting processes. Are there really three classes of quality control agents who, by virtue of using different technologies, are situated at different stages in the review process? In order to investigate this – and our subsequent studies which we present – we used the time to revert as a primary metric. The lower the time to revert, the less time a 'damaged' version of an article is visible to readers, and so Wikipedia researchers long have used the median and/or average time-to-revert to operationalize the community's quality control processes [5, 7, 10].

Time-to-revert is defined as amount of time between when a given edit is made to a page and when the edit is entirely removed from the encyclopedia in a later edit, taking the article back to a previous state. It is important to note that refining an edit is not considered a revert, and Wikipedians are discouraged from excessively reverting edits made in good faith. As such, reverts are the predominant mechanism used to remove undesirable material, notably vandalism and spam, from the encyclopedia. While we do not know the quality of the edits being reverted, reverts are for removing undesirable content, whether intentional vandalism, good-faith mistakes, or out of date material.

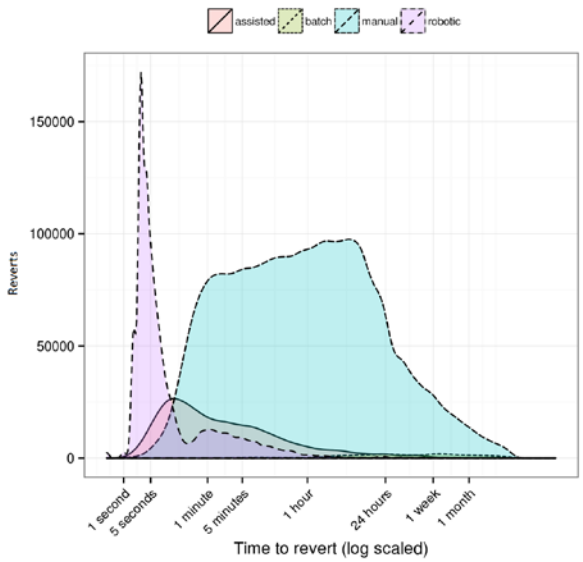

Figure 2: Histogram of time-to-revert during a normal month (January 2011), separated by different editing technologies.

2.3 Time-to-revert by type of revert tool used

In order to study the different types of reverting agents, we performed a series of analyses measuring the various distributions of time-to-revert for each reverting agent type. Figure 2, a histogram of time-to-revert by tool type, is based on data from all reverts made in during January 2011, when there were no major outages of countervandalism bots or tools in Wikipedia. For each reverted edit, we measured the time-to-revert and identified twelve different tools, bots, and browser-based actions commonly used for reverting edits, as well as unassisted, in-browser reverts. We distinguished between the different tools using the techniques of "trace ethnography" described in [3], mining edit summaries for standard traces that are automatically left by bots, tools, humans, and the MediaWiki software upon reverting edits.

We placed each of these editing technologies into the classes of reverting tools discussed in [4], then analyzed the time it took for members of each class to revert damaging edits. The time-to-revert for *manual reverts* – reverts that take place in the web browser and require that a user independently discover an edit to revert – occurs mostly between one minute and 24 hours. *Assisted reverts* – reverts made by human using tools that aid in identifying damaging edits and performing the revert – were substantially faster than humans editing manually, but performed far fewer reverts, overall. Fully-automated *robotic reverts* (e.g. ClueBot NG) are substantially faster. The overwhelming majority of robotic reverts are made within one minute the offending edit. Finally, we identified several automated agents that do not cleanly fit into these three categories. These are not bots continuously scanning for live edits like ClueBot NG, but instead are more often run as ad-hoc scripts. These *batch reverts* have a far more scattered distribution of time-to-revert and are not used to

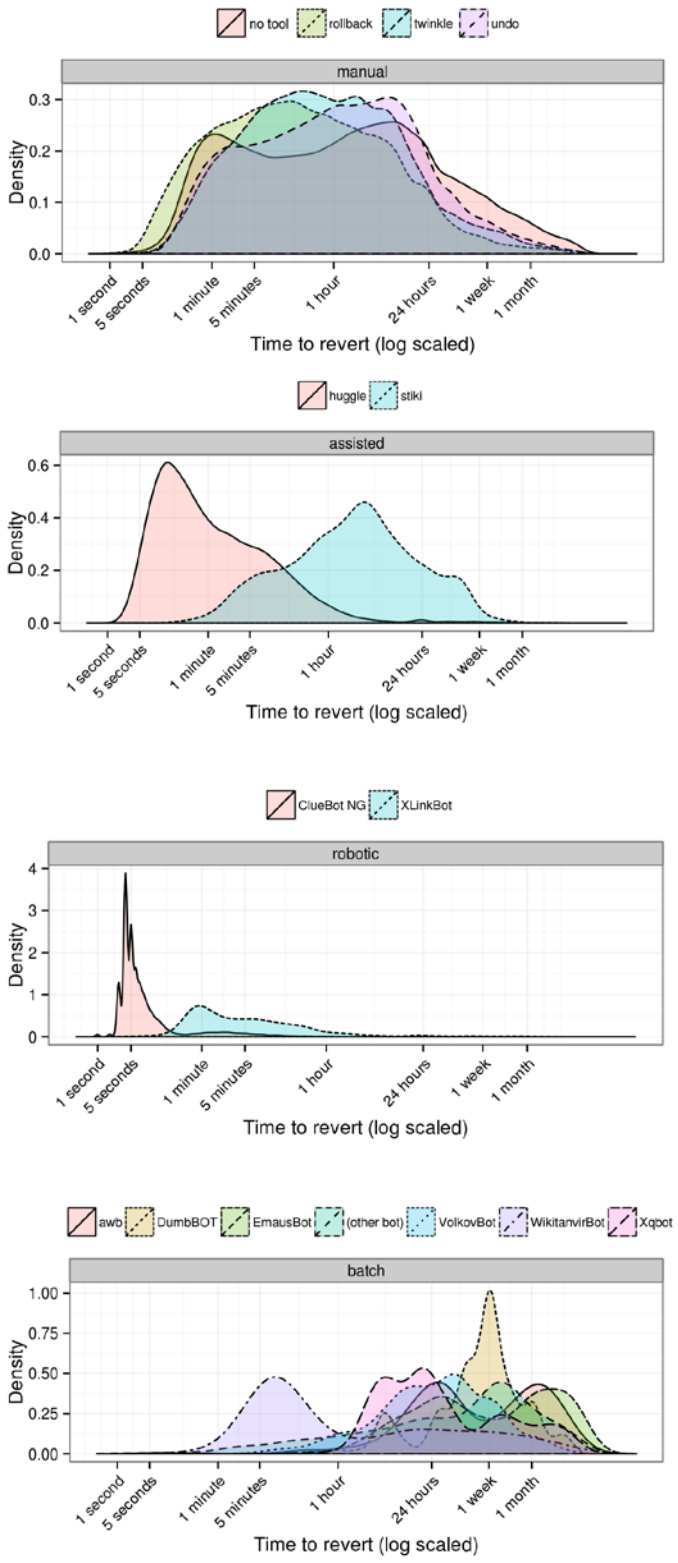

Figure 3: Time-to-revert probability density functions for individual revert tools (January 2011). The area under each curve represents the relative likelihood that a user using a particular tool will revert an edit during a given time period.

make nearly as many reverts as other tools, as they are used when, for example, Wikipedians decide to systematically change a category name or undo all edits by a malfunctioning bot.

2.4 The temporal rhythms of revert technologies

We next sought to distinguish between the individual tools used to make reverts in each of these four categories. We produced a set of empirical probability density functions representing the time it takes for editors using different technologies to revert any given edit (Figure 3). Note that these are not histograms representing the raw number of reverts, but instead the area under the curve represents the probability that a revert made with each individual tool will occur within a given period of time. By visualizing the activity of reverting agents this way, we can more easily compare the time-to-revert for Wikipedia's quality control processes, since some tools are used far more than others. Note that *batch* time-to-reverts are barely visible in Figure 2. This visualization is quite revealing when examining the tools *manual reverts*, for example. These include not only edits made with no assistance (by clicking the edit button, removing the offending text, and clicking "save"), but also in-browser functions built into MediaWiki, like rollback and undo, and JavaScript based user interface extensions like Twinkle. While these types of tools tools make it faster and easier to revert low quality edits, but like using the edit button, they still require the human editor to manually search for an offending revision to revert. As such, fully manual and partially assisted tool reverts have relatively similar distribution of time-to-revert probabilities.

This is in contrast with reverts made with fully assisted edit tools, (*assisted reverts*). These tools both make it faster and easier to revert low quality edits and give users a list of pre-screened edits to review in a queue. Most reverts with Huggle, the most widely-used, fully assisted, counter-vandalism tool, were made within 1 minute of the offending edit. It is interesting that reverts with STiki, a newer and more sophisticated queue-based vandal fighting tool, are more often made to somewhat older edits, with a time-to-revert distribution that is closer to unassisted edits. This suggests that Huggle and STiki are targeting different kinds of edits, likely of different levels of quality requiring different kinds of review.

When fully automated counter-vandalism bots revert an edit (*robotic reverts*), there is no human involved in either identifying or performing the revert. ClueBot NG and XLinkBot, the two main bots used to automatically revert edits during this time, perform reverts far faster than any of the other reverting agents. ClueBot NG is the fastest by a wide margin. Nearly every revert it makes is within 30 seconds of the original revision, and a substantial majority are within 5 seconds. XLinkBot is slower than ClueBot NG as well as most Huggle users, but faster than humans reverting manually and STiki users, with reverts about 1 minute of the offending edit.

We also saw many bots that were reverting edits at quite different periods of time than any other agents. This is because bots like AWB, DumbBOT, and EmausBot are less like the fully-automated counter-vandalism bots and closer to batch scripts that routinely perform cleanup, administrative, and categorization tasks. In one case, an editor made thousands of edits which were found to be malicious, all of their edits were systematically reverted using a batch script weeks later. While all of the edits we identified as reverts were technically restoring an article to a previous state, some of these actions may not be commonly considered to be "reverts" by editors. Bots like DumbBOT, which removes the standard article protection notices when a protected page has been unlocked by an administrator, is also technically analyzed as 'reverting' edits. Because unprotections routinely occur within one week of a page being protected, we see a noticeable spike at one week for DumbBOT. However, it is important to note that the raw number of "reverts" made by bots like DumbBOT is substantially lower than other agents.

3. THE BOT GOES DOWN

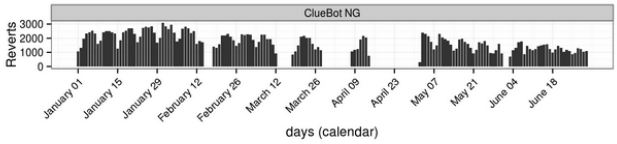

Figure 1: Edits by ClueBot NG from 1 Jan to 30 Jun, 2011

The next section uses the case of ClueBot NG's downtime as a natural experiment to investigate what the varying degrees of automation of quality control processes have afforded the Wikipedian community. As Figure 1 shows, there are many clearly defined periods of time in which this quality control agent was and was not operating, giving us a unique opportunity to examine what happens when such a prolific editor is suddenly no longer performing the duties it once did. We study this issue with two specific analyses, investigating differences between when ClueBot NG was and was not operating: 1) did it take Wikipedians longer to revert edits without ClueBot NG, and 2) were fewer edits reverted as a result of the downtime?

In order to answer both of these questions, we needed to find a way to cleanly compare the different kinds of time periods in which the bot was and was not operating. Wikipedia editors have quite different temporal rhythms [13], with daily and weekly cycles dramatically affecting how many edits are made and the quality of edits as well. In fact, West et al. found that these cycles can be quite useful in predicting future acts of vandalism [12]. Figure 1 also shows these weekly fluctuations. Simply comparing the number of reverts or even the proportion of reverts to edits made might be misleading due to these periodic

fluctuations. In order to control for this we needed to fairly compare time periods when ClueBot NG was down to when the bot was up. We found that most periods in which ClueBot NG was not operating included a Wednesday and a Thursday, so we limited our analysis to activities during those days of the week. We examined all Wednesdays and Thursdays from 1 January to 30 June 2011 and used various metrics to compare the differences between when the bot was and was not operating on these days.

3.1 Did it take longer to revert edits?

Using a similar approach to plotting time-to-reverts as in Figure 2, we produced a histogram of the time it took for different classes of quality control agents to revert edits. Figure 4 plots the total number of reverts made during the Wednesdays and Thursdays when ClueBot NG was down and up during the first half of 2011. It is important to note that the bot was up for far more days than it was down, so the raw number of reverts should not be taken as the significant factor in this plot. Rather, this plot shows that when the bot was down, there were simply no other agents that were reverting edits as fast as ClueBot NG had been. Assisted editors arrive on the scene first and revert some edits, followed by those performing reverts manually. Due to the smaller scale at which they operate, batch tools are barely noticeable when the bot was up or down. This suggests that when ClueBot NG stopped operating, we should expect to see an increase in the median time to revert time-to-revert. In our analysis, we found a substantial difference. As Table 1 shows, during ClueBot NG's downtime, the median time-to-revert nearly doubled.

3.2 Overall, were fewer edits reverted?

Having found that it took longer for Wikipedia's remaining quality control agents to revert edits when ClueBot NG was down, we wanted to understand whether or not this had an effect on the total number of edits that were reverted. If Wikipedians have a certain capacity for reviewing and reverting edits, and if ClueBot NG does a large portion of that work, Wikipedians may be unable to to make keep up with the workload when the bot went down.

We looked for evidence of such an effect by observing the difference between the proportion of new edits that were reverted different during ClueBot NG's downtime using our dataset of reverts occurring on Wednesdays and Thursdays during the first half of 2011. Figure 5 shows the proportion of reverted and reverting edits from this dataset. We found that the proportion of reverting edits was statistically significantly lower when ClueBot was down (χ^2 = 115.9, $p < 0.001$). However, the proportion of revisions that were eventually reverted was not significantly different (χ^2=0.64, p=0.43). In other words, while less reverting took place when ClueBot NG was down, the same proportion of revisions made during ClueBot NG's downtime were *eventually* reverted. This suggests that Wikipedia's quality control system is resilient in that all of the revisions that would have been reverted with ClueBot NG around will also be reverted when ClueBot NG is not around, but it will take longer -- long enough in this case that it was back online by the time that some of the reverting revisions took place (note that ClueBot NG did not perform these reverts and therefore did not make up for its own absence).

Table 1. Time-to-revert by ClueBot NG's status, Wednesdays and Thursdays only

ClueBot NG status	*Time-to-revert, Geometric mean*	*Time-to-revert, Median*
Up	941 sec (15.7 mins)	744 sec (12.4 mins)
Down	1674 sec (27.9 mins)	1286 sec (21.4 mins)

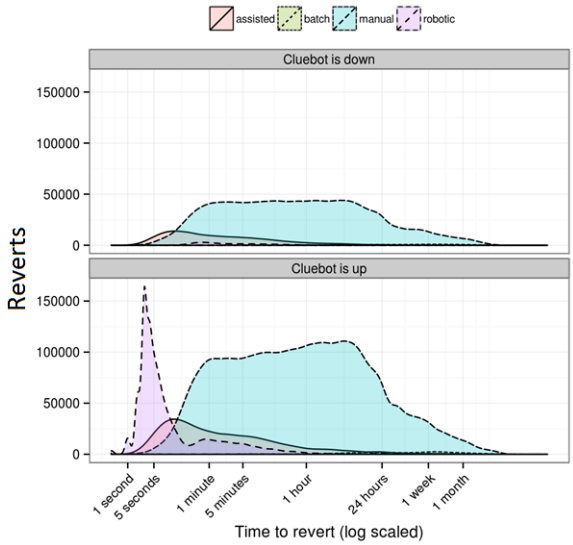

Figure 4: Histogram of time-to-revert during Wednesdays and Thursdays when ClueBot NG was and was not operating, separated by different editing technologies.

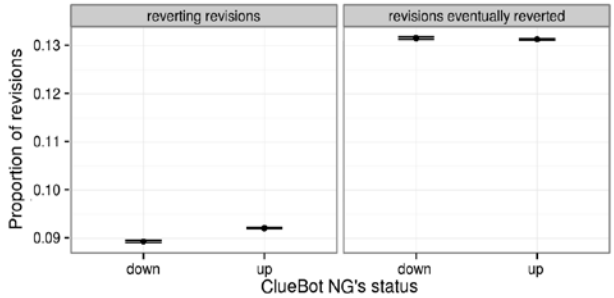

Figure 5: Proportion of revisions that were reverts and proportion of revisions that were eventually reverted.

4. CONCLUSION AND FUTURE WORK

This study has investigated the constitution of the heterogeneous quality control network in the English-language Wikipedia. A major contribution of this paper is in quantitatively modeling the temporal rhythms at play in this socio-technical system using an ethnographically-informed analysis of trace data. In this paper, we

conceptualize Wikipedia's quality control network as a multi-tiered system in which different classes of reviewers use various technologies to patrol for damage at different time periods. We found four types of reverting agents who act in different time scales: first, the near-instantaneous fully-automated robots, then rapid tool-assisted humans (cyborgs), followed by humans editing via web browsers, and finally, the more idiosyncratic batch scripts.

We can assume that the technologies used at each stage of the review process are used to examine different kinds of potentially-damaging edits. If we are correct in this assumption, automated bots revert the blatant offenders, assisted vandal fighters revert less obvious damage, and unassisted editors manually revert the more subtle damage, mistakes and norms violations. Future work may include edit quality as a factor to determine if this assumption holds true; if so, we can specifically examine how these editing tasks were redistributed. We might also distinguish between different kinds of reverts, as not all of the reverts we examined were necessarily removing vandalism.

Another major contribution of this study is in demonstrating the resilience of Wikipedia's decentralized quality control system when one of its core agents fails. ClueBot NG routinely reverts thousands of damaging edits every day, reverting most of these edits within five seconds. The overall time it took for any given Wikipedian to revert an edit almost doubled when this software agent went offline. While a significantly smaller number of quality control took place during the bot's downtime, we found that a similar proportion of were eventually reverted. This suggests that other agents in Wikipedia took over this quality control work, but performed it at a slower rate.

There are many unanswered questions which we leave for future study. First, we do not wish to imply that ClueBot NG is not a critical actor because Wikipedia's quality control network was resilient and redistributed tasks when it went down. An increase in the median time-to-reverting low quality edits from 12.4 minutes to 21.4 is substantial considering the many roles that Wikipedia plays in the global information ecosystem. Another important concern is the continued viability of such a workaround: tasks were redistributed from bots to humans, but at what cost? The last major stretch of downtime in this period was two weeks, and if our title is indeed correct in framing this as a kind of natural disaster, that is a long period of time for editors to be in an exceptional, workaround mode, performing tasks that they would not usually perform. When Wikipedians made up for the loss of ClueBot NG, did they have to sacrifice the time they may have spent constructing the encyclopedia to protect it from vandalism? This is a question which can be answered in a qualitative study of the reactions that Wikipedians had to ClueBot NG's downtime. Given the scope of this study, we have omitted a substantial amount of context. How did Wikipedians react to ClueBot NG's downtime? Did vandal fighters even notice; if they did, were their responses coordinated or emergent? Are similar processes at work in other language versions of Wikipedia and other wikis?

Furthermore, who were these editors who 'took up the slack'? Were they dedicated vandal fighters using tools like Huggle and STiki, or were they editors who generally performed other tasks in the encyclopedia project? Could some of them be newcomers who were drawn to edit by reverting vandalism? We also make several assumptions about the quality (or lack thereof) of edits reverted based on the time it takes to revert an edit. By examining the quality of individual edits using hand-coding or algorithmic metrics, we can better understand how the distribution of work shifted and at what cost to the system.

5. ACKNOWLEDGMENTS

This work would not have been possible without the support of our research groups, the Wikimedia Foundation, NSF grants IIS 09-68483 and IIS 11-11201, as well as the constructive comments from the WikiSym reviewers.

6. REFERENCES

[1] Bowker, G.C. and Star, S.L. 2000. *Sorting Things Out: Classification and Its Consequences*. MIT Press.

[2] Geiger, R.S. and Ribes, D. 2010. The work of sustaining order in wikipedia: the banning of a vandal. *Proc CSCW 2010*.

[3] Geiger, R.S. and Ribes, D. 2011. Trace Ethnography: Following Coordination Through Documentary Practices. *Proc HICSS 2011*.

[4] Halfaker, A. and Riedl, J. 2012. Bots and Cyborgs: Wikipedia's Immune System. *Computer*. 45, 3, 79–82.

[5] Kittur, A., Chi, E., Pendleton, B.A., Suh, B. and Mytkowicz, T. 2007. Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. *Proc alt.CHI 2007*.

[6] Kriplean, T., Beschastnikh, I. and McDonald, D. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. *Proc CSCW 2008*, 47–56.

[7] Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L. and Riedl, J. 2007. Creating, destroying, and restoring value in wikipedia. *Proc GROUP 2007*, 259–268.

[8] Star, S.L. 1999. The Ethnography of Infrastructure. *American Behavioral Scientist*. 43, 3, 377–391.

[9] Stvilia, B., Twidale, M.B., Gasser, L. and Smith, L.C. 2008. Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*. 59, 6, 983–1001.

[10] Viegas, F., Wattenberg, M., Kriss, J. and Van Ham, F. 2007. Talk Before You Type: Coordination in Wikipedia. *Proc HICSS 2007*.

[11] West, A.G., Kanna, S. and Lee, I. 2010. STiki : An Anti-Vandalism Tool for Wikipedia using Spatio-Temporal Analysis of Revision Metadata. *Wikisym 2010*, 2–4.

[12] West, A.G., Kannan, S. and Lee, I. 2010. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? *Proc EUROSEC 10*. 1752050, 6, 22–28.

[13] Yasseri, T., Sumi, R. and Kerétsz, J. 2011. Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLoS ONE*. 7, 1, e30091.