**Wikipedia edit source analysis: Cultural implications**

The culture of ODL

**Abstract**

This study determined what percent of the edits to different types of Wikipedia articles were made by various types of users. The specific article types studied were Wikipedia’s featured articles, its 1,000 most frequently visited articles, the articles mentioned on the Wikipedia subreddit on Reddit.com, and randomly selected articles. The user types studied were automated users, administrators, frequent editors, infrequent editors that are nevertheless registered, and unregistered users. All article and user category types were mutually exclusive.

Across all article categories, infrequent and unregistered users made the most edits. However, when edits were weighed for content, frequent editors contributed solid majorities of the edits to featured and randomly selected articles and, depending on the threshold for being considered a frequent editor, contributed pluralities of weighed edits to the most popular articles and articles posted to the Wikipedia subreddit. Overall, 0.29% of the observed editors contributed 88% of the weighed content. Majorities of the weighed edits to the various article types were made by different edit ranges of frequent editors. For example, the weighed edits to featured articles were almost entirely contributed by the 100 most frequent editors whereas the weighed edits to articles posted on the Wikipedia subreddit were primarily made by the 3,001-3,500 most frequent editors.

That different ranges of frequent editors reliably contribute to different types of articles implies that those using Wikipedia as a means of researching niche topics, reading about current events and pop culture, or seeking entertainment will likely see content generated by different ranges of frequent editors. Also, this study observed that frequent editors tend to make much more substantial edits than infrequent editors. If Wikipedia wants its content to be written by a wider base of users, it will likely have to make changes, such as its recent simplification of its editing interface. There are other changes it could enact, such as changing its culture regarding article deletion. Finally, that so much of Wikipedia’s content is generated by such a small group of users suggests that other educational online communities may see similar patterns.

**Keywords**: Wikipedia Edits; Wikipedia Culture; Frequent Editors; Webscraping; Scrapy

**Introduction**

Wikipedia is an online collaborative encyclopedia. Both the writing and reading of it have been studied as educational tools.[[1]](#endnote-2)[[2]](#endnote-3)[[3]](#endnote-4)[[4]](#endnote-5) It is also read for entertainment.[[5]](#endnote-6)[[6]](#endnote-7) Since most articles on Wikipedia are editable by anyone, there have been inquiries into what sort of user actually writes most of its content.[[7]](#endnote-8)[[8]](#endnote-9) Is most of the content generated by an elite group of a few hundred users that all know each other as the founder of Wikipedia Jimmy Wales has claimed or do the masses generate most of the content through their knowledge about niche topics?[[9]](#endnote-10) To what extent do editors get edits to their name by making superficial changes to the content generated by those who have made far fewer edits?

If Wikipedia is indeed written by a relatively small group of users, then perhaps that is because the culture of Wikipedia discourages newcomers. Specifically, there have been criticisms of a culture of “deletionism” in which the content of newcomers is deleted for minor reasons, thus discouraging them from making further contributions.[[10]](#endnote-11)[[11]](#endnote-12) Also, while the governing structure of Wikipedia has been found to be egalitarian,[[12]](#endnote-13)[[13]](#endnote-14) that egalitarianism becomes less significant if potential users are dissuaded from participating. Furthermore, because Wikipedia is one of the largest and most successful online educational communities,7 illuminating its culture would provide an expectation as to what sort of culture could be expected in other such communities. Specifically, does Wikipedia followed the Pareto Principle in which case 80% of the content would be generated by 20% of the users?13 If it does, that would be reason to expect similar communities would as well.

**Previous Research**

A few studies have sought to answer this question. One manually analyzed 250 recent edits and found that only 32 of those edits added substantial content to their respective articles.[[14]](#endnote-15) None of those edits were contributed by administrators. Also, none of the users that made those edits had a received a "barnstar," an ad hoc award given in thanks to helpful editors. Furthermore, 31% of the users who made a substantial edit were not even registered. These results indicate that much of the content on Wikipedia is generated by infrequent contributors. Another study looked not just at the user type of each editor, but also at the number of edits that that user had made.[[15]](#endnote-16) It found that registered users' edits increased in quality with their number of edits, but that unregistered users’ edits decreased in quality. That result implies that users should be classified not just by their type, i.e. whether or not they are an administrator or an unregistered user, etc, but also into subtypes based on their number of edits. Lastly, one study found that Wikipedia was initially written by a relatively small group of very frequent editors, but that it is now written by the masses to a much greater extent.

**How This Study is Different**

All of those studies implicitly chose to focus on randomly selected articles. That is logical in the context of scientific investigations’ common emphasis on random sampling. However, just analyzing random articles is not representative of the articles people typically read as some articles are read much more frequently than others. Still, it also would not be optimally representative to just look at the most popular articles because Wikipedia is also used to research niche topics and for entertainment.5 As such, this study differs from those previous to it in that it looks at different types of articles and seeks to determine who generates their content. Specifically, it analyzes the most frequently visited articles on Wikipedia, randomly selected articles, articles of such quality that Wikipedia has designated them as featured, and articles interesting enough to be mentioned on the Wikipedia subreddit on Reddit.com. Reddit is a website through which users submit and vote on content and "subreddits" are communities on Reddit devoted to various things people are interested in.[[16]](#endnote-17) It is hoped that articles posted to Reddit’s Wikipedia subreddit will be representative of the sort of content that people who use Wikipedia as entertainment will browse. Similarly, it is hoped that the randomly selected articles will be representative of niche topics.

**Data Collection**

The overall data collection process of this study involved scraping lists of each type of user, lists of articles from various sources, and the edit history of each article. That process was automated by using the python webscraping library Scrapy. Such automation was possible because the data is kept in consistent locations specified in the various web pages’ XML. Ultimately, the histories of 1,000 randomly selected articles, 730 articles posted to the Wikipedia subreddit, 100 featured articles, and the 100 most popular articles in June 2013 were scraped. This study's scraping and analysis code, an explanation of that code, and a list of the relevant xpaths for those that want to create their own code is available online.[[17]](#endnote-18)

**User Group Divisions**

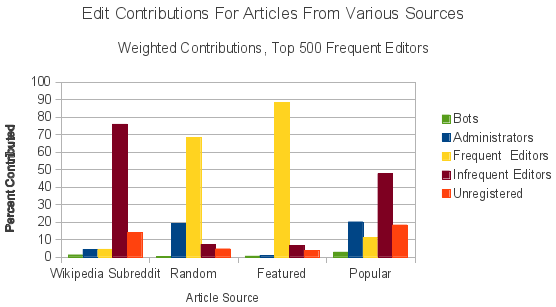
Ultimately, each edit is interpreted as being contributed by an automated user or “bot,” an administrator, a frequent editor, an infrequent but registered editor, or an unregistered user. Such classification requires creating discrete categories of users despite the fact that many users span multiple categories. For example, there are many bots that are administrators. Also, there are many administrators that are frequent editors. That issue is addressed by classifying anything that is a bot as a bot and anything that is an administrator, but not a bot, as an administrator. Any user that is a frequent editor that is neither an administrator nor a bot is classified as a frequent editor. Registered users that do not fit into any of the above categories are classified as infrequent editors and anonymous editors are classified as unregistered users.

**Edit Weighing Process**

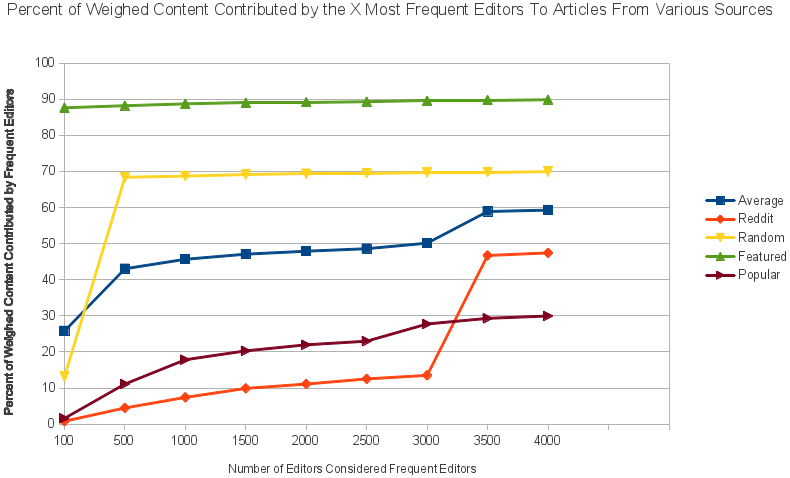
The edits contributed by each user can be interpreted in two ways. One way is to view each edit as equivalent and simply count them. That entails keeping a tally of the amount of edits contributed to each article type by each user type. Another way is to attempt to weigh the edits by their content so that the magnitude of each edit affects how much it influences the editing user type’s tally for a given article type. In this study, when edits add content to an article, every additional character increases the weight of the edit by one. For negative contributions, however, the weight only increases with the magnitude of the edit until the magnitude reaches 10. At that point, all further characters removed do not increase the weight of the edit. This cap of 10 on the weight of negative edits is imposed because determining an entire paragraph is of poor quality most likely does not take linearly more effort than determining a sentence is of poor quality. It may even be easier to just delete a paragraph than to identify its salvageable sentences.

**Results**

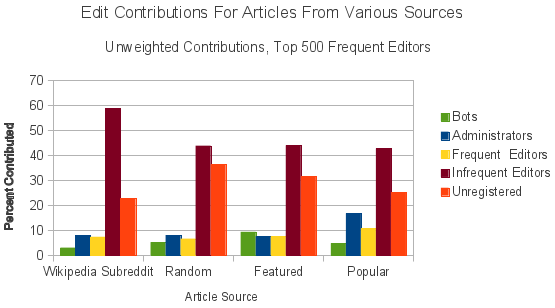
When considering the results of this study, it is important to keep in mind that the division between frequent editors and infrequent editors can be arbitrarily set. Thus when discussing the percentages of content contributed by one of those user groups, the threshold for being considered a frequent user, whether one must be in the top 100 or even in the top 4,000 most frequent editors, will always be specified. Also, two-way chi square contingency tests found that the probability that the number of edits and weighed content made is independent of user and article type to be less than 0.0001.



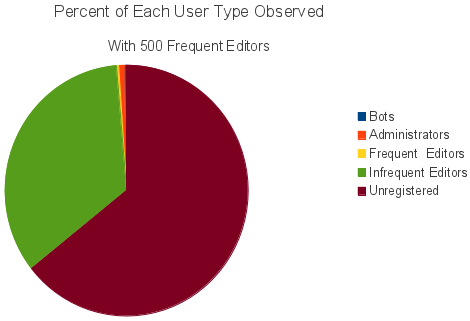
When at least 500 users are considered frequent editors, that category is the most significant contributor of weighed content to both randomly selected articles and featured articles. Specifically, 68.35% of the weighed content contributed to the randomly selected articles came from frequent editors. That is more than three times the 19.3% contributed by the second highest contributor to that category, administrators. The dominance of frequent editors over the featured article category is even more impressive. 88.17% percent of the content contributed to featured articles came from frequent editors, which is over 13 times greater than the 6.62% contributed by infrequent editors. Articles posted to the Wikipedia subreddit and the most popular articles, however, are written by a plurality and a majority of infrequent users respectively. When the 4,000 most frequent editors are considered frequent editors, they gain a plurality of the contributions to articles that were posted to the Wikipedia subreddit and the most popular articles.

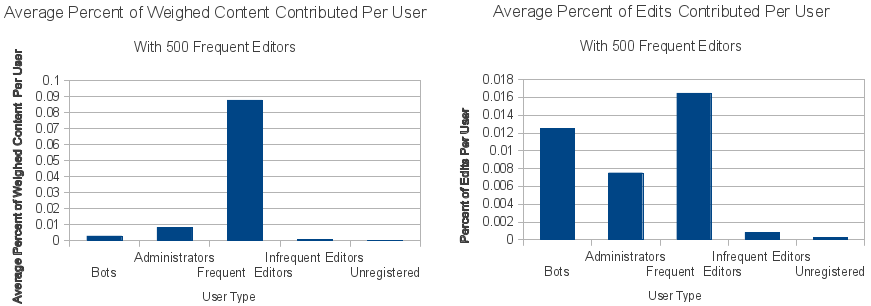


As all article categories are dominated by either frequent or infrequent users, as opposed to bots, administrators, or unregistered users, it is informative to look at how the percent of weighed content contributed by frequent editors varies as the number of users considered frequent editors is increased. Doing so reveals that frequent editors cluster together in the sort of articles they edit by their edit frequency. Most strikingly, 87.59% of the weighed content contributed to featured articles are contributed by the 100 most frequent editors. When looking at the content produced by the 4,000 most frequent editors, that figure only increases by 2.55% to 89.82%. Conversely, when moving from the top 100 frequent editors to the top 500, the share of content contributed to randomly selected articles increases by 418.59% from 13.18% to 68.35%. There is similarly a 241.49% jump when looking at the top 3,500 most frequent editors’ contributions to Reddit articles rather than the top 3,000. Furthermore, much more content is contributed by the 101-500 most frequent editors to randomly selected articles than the first 100 and the 500-4,000 most frequent editors combined. These figures indicate that while much of Wikipedia is written by a small group of frequent editors, different types of articles are written by subdivisions of that small group based upon edit frequency. Were someone to visit an article after seeing it on the Wikipedia subreddit as opposed to seeing it on Wikipedia’s list of featured articles, they would be looking at content that was contributed by a very different group of editors - the 3,001 to 3,500 most frequent editors and the 100 most frequent editors respectively. This finding is reminiscent of previous research that found it is important to look at the number of edits a user has made in addition to their user type to predict edit quality.



Despite the frequent editor group’s dominance over weighed edits, infrequent and unregistered users made the largest total number of edits to articles of all types. That is, except when at least 3,500 users are considered frequent contributors as at that point the plurality is gained by infrequent and frequent users. That such a large amount of edits came from infrequent and unregistered users is to be expected as far more users fall into one of those groups than any other user group. What is surprising is how much the share of frequent editor contributions fall when edits are not weighed by magnitude. That result indicates that edits made by frequent users tend to be more substantial than those made by unregistered and infrequent editors, which contradicts the theory that infrequent and unregistered users produce most of the real content on Wikipedia while frequent editors mainly make organizational and cosmetic edits.





Furthermore, on a per user basis, the average percent of weighed content contributed by frequent editors dwarfs those of other user types. The average frequent editor, when 500 users are considered frequent editors, contributed more than 10 times as much weighed content as administrators and more than 30 times as much weighed content as automated users, the second and third greatest contributors of weighed content respectively. However, frequent users only made 31.4% more edits than bots and a little over double the number of edits made by administrators. That the edits bots made were so poorly weighed validates the weighing mechanism to some extent in that they would be expected to make many unsubstantial edits.

**Cultural Implications**

That frequent editors provide so much of the weighed content to Wikipedia does suggest that Wikipedia is written more by elite users than by users with knowledge about a niche topic. Still, the elite frequent users seem to be divided into subclasses in which users with different numbers of edits modify different types of articles. This implies that despite Wikipedia’s apparent uniformity, the articles a person sees will be written by varying classes of editors depending on what that person primarily uses Wikipedia for.

The concentration of edits among a relatively small group of users is potentially indicative of the sort of participation that exists in other educational online communities. Although only 0.29% of the users observed making an edit were one of the 1,000 most frequent editors, 88% percent of the average weighed content across all categories came from them. The Pareto Principle holds that 80% of the results will be caused by 20% of the causes.13 Wikipedia seems to follow a much more extreme version of that rule. This suggests that other online educational communities will as well.

With that said, it is possible that the dominance of a relatively small group of users over Wikipedia’s weighed edits stems from issues specific to Wikipedia. Perhaps the article editing system is too complex for the inexperienced. A simplified version of the editing system was released recently, which may increase non-elite participation in content creation.[[18]](#endnote-19) Wikipedia’s culture may also discourage newcomers if it deletes their content because it did not meet various standards. If Wikipedia would prefer for its content to be produced by a larger spread of users, it should research such barriers to entry.

**Potential Sources of Error**

This study’s weighing mechanism naively assumes that all contributed content is of equal quality. However, many edits are vandalism. Also, even content that is not vandalism gets removed. The amount of persistent content contributed is probably more important than the size and number of a user type’s edits.

The issue of vandalism could be addressed without drastically changing the methodology of this study by identifying any edit as reverted vandalism if it is followed shortly by another edit of the same size but with an opposite sign. The precision of such a heuristic could be improved by requiring that the opposite-signed edit be contributed by a bot as there are bots set up specifically to detect and revert vandalism.[[19]](#endnote-20)

Also, the quality of a contribution is determined by many factors other than just word count. One method to get around that is manually scoring each individual edit. Alternatively, a more complex weighting heuristic could use factors like citations as an indicator for quality and the presence of vulgar words as a signal for vandalism.

That different numbers of the various article types were analyzed is mostly insignificant because user edit percentages were studied by article type. However, statistics involving all observed edits were influenced by how many of each type of article was analyzed as different types of users contribute to different types of articles. With that said, featured and popular articles tend to have much longer edit histories than the other article types so analyzing fewer of them makes the overall statistics more representative of each type of article.

Also, the classification of all IP addresses as unregistered users is flawed in a few ways. It is possible that an anonymous edit was actually made by a frequent editor that forgot to log in. Also, IP addresses do not necessarily represent just one user. It is possible that a single IP address could represent an entire school for example. This study therefore may have underestimated or overestimated the number of unregistered users that contribute edits to Wikipedia.

**Potential For Further Research**

It is very likely that the share of edits made to an article by different user groups varies not just with article source, but also with time. That is, perhaps the users who contribute to an article when it is just beginning to be written vary significantly in type from those who only begin editing it once it is up for nomination as a featured article. If it is assumed that an article's first editors contribute most of its content and that the latter editors mostly improve the presentation of that content, then it would be logical to weigh edits for substantiality by time rather than just edit magnitude.

**References and Notes**

1. Chandler, C. J., & Gregory, A. S. (2010). Sleeping with the Enemy: Wikipedia in the College Classroom. History Teacher, 43(2), 247-257. [↑](#endnote-ref-2)
2. Bravo, V. J., & Young, M. F. (2011). The Impact of a Collaborative Wikipedia Assignment on Teaching, Learning, and Student Perceptions in a Teacher Education Program. Canadian Journal Of Learning And Technology, 37(3). [↑](#endnote-ref-3)
3. Sormunen, E., & Lehtio, L. (2011). Authoring Wikipedia Articles as an Information Literacy Assignment: Copy-Pasting or Expressing New Understanding in One's Own Words?. Information Research: An International Electronic Journal, 16(4). [↑](#endnote-ref-4)
4. Cattapan, A. (2012). (Re)Writing "Feminism in Canada": Wikipedia in the Feminist Classroom. Feminist Teacher,22, 2 ,125-136. [↑](#endnote-ref-5)
5. Lim, S. (2009). How and why do college students use Wikipedia?. Journal Of The American Society For Information Science & Technology, 60(11), 2189. [↑](#endnote-ref-6)
6. Lim, S., & Kwon, N. (n.d). Gender differences in information behavior concerning Wikipedia, an unorthodox information source?. Library & Information Science Research, 32(3), 212-220. [↑](#endnote-ref-7)
7. Townsend, S., Osmond, G., & Phillips, M. G. (2013). Wicked Wikipedia? Communities of Practice, the Production of Knowledge and Australian Sport History. International Journal Of The History Of Sport, 30(5), 545-559. [↑](#endnote-ref-8)
8. Swartz, A. (2006). Who Writes Wikipedia?. Retrieved from http://www.aaronsw.com/2002/whowriteswikipedia/. [↑](#endnote-ref-9)
9. Kittur, A.,Chi, E. H., Pendleton, B. A., Suh, B., Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. Alt.CHI. [↑](#endnote-ref-10)
10. Stvilia, B., Twidale, M. B., Smith, L. C. and Gasser, L. (2008), Information quality work organization in wikipedia. J. Am. Soc. Inf. Sci., 59: 983–1001. doi: 10.1002/asi.20813. [↑](#endnote-ref-11)
11. Halfaker A., Kittur, A., & Riedl,J. (2011). Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In Proceedings of the 7th International Symposium on Wikis and Open Collaboration. WikiSym '11. ACM, New York, NY, USA, 163-172. DOI=10.1145/2038558.2038585 http://doi.acm.org/10.1145/2038558.2038585 [↑](#endnote-ref-12)
12. Konieczny, P. (2009). Governance, Organization, and Democracy on the Internet: The Iron Law and the Evolution of Wikipedia. Sociological Forum, Vol. 24, No. 1, pp. 162-192 [↑](#endnote-ref-13)
13. Spek, S., Postma, E., & Van den Herik, J. (2012). Wikipedia: Organization from a Bottom-Up Approach. Research in Wikipedia, on the Wikisym 2006. Retrieved from http://ssrn.com/abstract=2097682. [↑](#endnote-ref-14)
14. Anthony, S. (2006). Contribution Patterns Among Active Wikipedians: Finding and Keeping Content Creators. Wikimania 2006 - Wiki Community Dynamics [↑](#endnote-ref-15)
15. Anthony, D., Smith, S. & Williamson, T. (2007). The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia. Dartmouth Computer Science Technical Report TR2007-606. [↑](#endnote-ref-16)
16. Reddit (2013). faq - reddit.com. Retrieved from http://www.reddit.com/wiki/faq [↑](#endnote-ref-17)
17. Weinstein, D. (2013). https://www.dropbox.com/sh/lwqdhavh7lssv5k/OG2LJ5UL5A/Wikipedia. [↑](#endnote-ref-18)
18. Sawers, P. (2013, June 6). Wikimedia Unwraps WYSIWYG Visual Editor for Logged-in Users. Retrieved from http://www.techzone360.com/topics/techzone/articles/2013/07/10/345241-wikimedia-unwraps-wysiwyg-visual-editor-logged-users.htm [↑](#endnote-ref-19)
19. Mola-Velasco, S. M. (2012). Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals: Lab Report for PAN at CLEF 2010. [↑](#endnote-ref-20)