Vandalism on Collaborative Web Communities: An Exploration of Editorial Behaviour in Wikipedia

Abdulwhab Alkharashi
University of Glasgow
a.alkharashi.1@research.gla.ac.uk

Joemon Jose
University of Glasgow
Joemon.Jose@glasgow.ac.uk

ABSTRACT

Modern online discussion communities allow people to contribute, sometimes anonymously. Such flexibility sometimes threatens the reputation and reliability of community-owned resources. Such flexibility is understandable, however, they engender threats to the reputation and reliability in collective goods. Since not a lot of previous work addressed these issues it is important to study the aforementioned issues to build an innate understanding of recent ongoing vandalism of Wikipedia pages and ways to preventing those.

In this study, we consider the type of activity that the anonymous users carry out on Wikipedia and also contemplate how others react to their activities. In particular, we want to study vandalism of Wikipedia pages and ways of preventing this kind of activity. Our preliminary analysis reveals (∼ 90%) of the vandalism or foul edits are done by unregistered users in Wikipedia due to nature of openness. The community reaction seemed to be immediate: most vandalisms were reverted within five minutes on an average. Further analysis shed light on the tolerance of Wikipedia community, reliability of anonymous users revisions and feasibility of early prediction of vandalism.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; *Vandalism*; • **Human-centered computing** → Wikis;

KEYWORDS

Crowdsourcing; Vandalism; Encyclopedia

1 INTRODUCTION

Web-based communities offer attractive problems to study the nature and dynamics of such collaborative systems. They provide the opportunity of empirically studying multiple aspects of user behaviour at a large-scale through the extraction of online datasets via dedicated tools, such as application programming interfaces. The most widely active collaborative platforms (i.e., Wikipedia, Reddit or Twitter) can be taken as an example of an online discussion community, where users of such platforms can freely participate in discussion and content production, establish links to other members and create and maintain affiliations to the variety of communities that these platforms support. The growth of data in terms of edits, comments or votes makes the analysis of vandalism behaviour a challenging task. In particular, it requires exhaustive study of massive dataset to unearth the association of different kind of user with type of edits and quality of edits [2]. We want to analyse the edit history [17] of different Wikipedia articles and find out interesting facts that can be extracted from it i.e., types of editors/contributors, recognizing editor characteristics from edits they make [3, 8], how do articles evolve, community reactions to edits and so forth. Previous work addressed different aspects of Wikipedia issues from editor behavior, influences, personality traits perspectives [2, 10, 15] to vandalism detection model [1, 6, 14]. Another study [12] shows how diversity influenced crowd performance under different conditions of task conflict and communication in Wikipedia article production. The idea of sharing knowledge collaboratively became so active, yet anti-social behaviour still exists for a number of reasons including stress or boredom [7] , which affects the total contribution and reliability of the wikipedia (WP). To prevent this, we need to identify what kind of editors are involved in the vandalism category and which types of articles are susceptible to these malicious activities. Thus, we form the following research questions:

***RQ1**: Is there an edit gap between anonymous and registered wikipedians in terms of the number of contributions?*

***RQ2**: How anonymous and registered users behave in the community based on categories of an article?*

***RQ3:** How contributors are reacting to vandalized edits over time?*

We seek to understand the vandalism behaviour of the Wikipedians to uncover the social patterns that resulted in such behaviour. This work addressed various challenges for making the study possible. First, the dataset of Wikipedia dump is of titanic scale. The English wiki dump produces terabytes of edit history which was cumbersome to handle. Second, the results of analysis are often contradicting and it is difficult to determine a comprehensive metric for the study. Third, due to the nature and size of data the presence of bias in study is difficult to monitor. The main contributions of this study is as follows:

- Identify vandals and the reliability of their contribution.
- Evaluate community reaction in terms of reversion made over edits.
- Study the editorial behaviour of anonymous and registered in Wikipedia.

The remainder of this paper describes these contributions in detail. To answer each research question, we describe how the data was collected from Wikipedia then report the evaluation of results before we finalize our conclusive remarks.

CERI'18, June 25-27 , Zaragoza, Spain
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6543-7…$15.00
DOI: 10.1145/3230599.3230608

2 BACKGROUND

The definition of vandalism would include multiple aspects from psychological point of view. Harriet et. al [4] defines vandalism as voluntary degradation that qualify behaviour and classify conditions when damages are intentionally targeted to an object. In collaborative web communities, vandalism behaviour refers to an edit that is offensive, deceptive and destructive in altering a content. In Wikipedia, contributors assess the vandalised edit to see if it was made in a good or bad faith manner before reverting an edit [11]. A good-faith revert in this case is when contributors edit to represent an opinion and is not malicious.

The purpose of this study is to identify anonymous user behavioural patterns in Wikipedia edits and corresponding community feedback. Wikipedia is the best online encyclopedia available which is free and open to edit for all. It is relatively simple to make an edit on an existing article. Registered users as well as anonymous users can modify any article any time using their IP address. Wikipedia stores all the edit history and the entire snapshot of the article is saved. Each article has a talk page and discussion page. People can talk and discuss before and after making an edit. Every registered editor has his own page where he can update his personal information and interest. Every user page also associate with a talk page which reflects the topics he is interested in. Using distance metric between two consecutive versions of articles, we can analyse how much change is made, the type of edits and type of editors. Further analysis might reveal the characteristics of the editors, which can be used to take preventive measures against foul edits/vandalism. It is significant to address the vandalism behaviour problem in the context of user role and the community feedback.

2.1 Understanding anonymous user behaviour

In order to better understand the behaviour of anonymous users, it is important to analyse the percentage of articles when they contribute. Later we focused on several questions and will try to answer them such as:

- How do anonymous users contribute? For example, if they prefer to stay on topic or they stray. So we need to compare the similarity between the current and previous edits made by anonymous users. We also have to measure the type of edits that they are making, for example, when they contain negative emotions/words.
- How do they gather popularity? For example, if they are contributing to a controversial topic and making further controversial edits, how community is reacting to that in talks or discussion pages. What is the number of threads in the relevant discussions initiated by the anonymous users?
- How anonymous users react to community reaction? We seek to know whether anonymous users remain silent or attempt to make further edits and what is the consequence.

2.2 Measuring community reaction

Community feedback is the key factor of this study. We aim to see whether the community is particularly harsh, flexible or lenient towards anonymous users. To measure this, we can count the number or percentage of times when their edits were reverted. It might turn out that these edits are more reliable and are less likely to be reverted. So it is not hard to judge from this metric what is the ratio of reverts for each user type. A deeper analysis might reveal answers if there is any community bias towards the registered users and if the tolerance of community towards anonymous users edits are evolving over time.

Table 1: Summary of wiki dataset

	#Reversions	#Registered	#Anonymous
Dataset A	158,148	2,795	5,779
Dataset B	148,317	2,590	5,987

3 DATASET

The data is collected from wikipedia [1] which gets updated each month and saved each year into a decentralized linked data system. Each dataset contains wiki pages from different time intervals in 2016 and 2017. The dumps for different language versions of Wikipedia pages are kept separate. It is important to note that the data are embedded in XML format and needed to be transformed into more readable format. We used a Java program, *mwdumper*, to perform the XML-SQL translation. Considering the huge size of English wiki dump, *mwdumper* is an efficient program that generates the script without getting crashed. We also used Wiki Edit History Analyzer which processes MediaWiki revision history and produces summaries of edit actions performed. Basic edit actions include insert, delete, replace, and move; high-level edit actions include spelling correction, wikify, etc. Data visualization is another important aspect of this study to get an insight of the edit history. Several interesting tools were used in different part of the study include: HistoryFLow, Listen to Wikipedia, StatMediaWiki, Wiki Explorer.

Both datasets were divided into 50 different tables under a predefined schema. The three main tables are user, page and revision. The remaining tables describe the page/user categories and their relations. table page/category gives information about page and categories. Similarly, table category provides information about user categories. Table 1 summarizes the selected dataset. The Reliability of user edit is measured by retention rate of articles given by,

$$R = \frac{\#character\_retained}{\#character} \times 100\% \tag{1}$$

The retention rate [3] is to determine the number of retained characters in an article for each user, divided by the total number of characters in each edited article by a user. It might be the case such edits can be reliable. Yet if not, then it can act as a precursor to preventing foul edits and vandalism. Since this study is preliminary, we attempt to provide a comprehensive overview of what most editors do in WP. In particular, explore editorial behaviour between anonymous and registered users in terms of edit activity, category of an article and community reaction.

4 RESULTS

This section is devoted to the findings of the analysis performed on the used dataset. The results are depicted using graphs. The analysis

[1] https://dumps.wikimedia.org

can be grouped according to different datasets and corresponding research questions. Following subsections, we attempt to elucidate the results.

4.1 Edit activity (RQ1)

We analyse the edit activity based on type of user who is targeting specific articles and compare the results with the data of most vandalized pages in Wikipedia. The result is depicted in Figure 1 describes the edit gap between registered and anonymous contributors. In particularly, anonymous editors tend to perform minor edits than registered editors and they do less in major edits. There could be variety of reasons that allow this phenomenon to occur including self-confidence, fear of revenge or other social reactions. For example, message boards established outside online communities, but for users of such community to vent their opinions on the user, have sometimes been used in ways that at least the communities themselves were not supportive about it [13], or for privacy and security related concerns [5, 9].

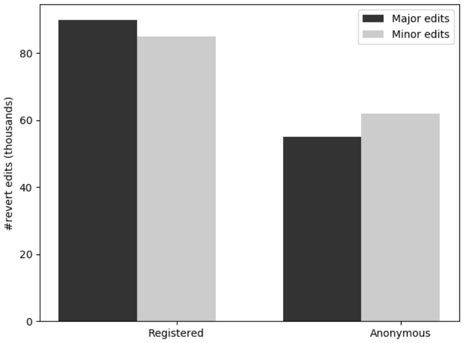

Figure 1: Edit activity of anonymous versus registered users by number of edit changes/reverts

4.2 Category of article (RQ2)

Vandalisms are often caused by lack of knowledge, attention seeking attitude, personal grudge and so on. It is important to understand not every malicious act is considered as vandalism. For example, abuse of tags, illegitimate page creation, spam external linking, trolling etc. are considered few of different types of vandalisms. The vandalism study was performed on the dataset for randomly selected articles. On an average, ∼ 90% of times the vandalism is caused by anonymous users. However, study on user pages yielded interesting result. Out of 10 randomly generated user page, the ratio (% of vandalism done by registered to anonymous) returned was 47:53. This might be indicative of the fact that, anonymous user tend to target main article pages while registered users are main culprit for vandalisms in user pages. The majority of the articles targeted by registered or anonymous users are related to Politics (29.4%), followed by Culture (26.4%), Music (23.5%) , Animals (11.7%) and History (8.8%). Intuitively the targeted articles for unregistered users tend to be controversial topics. The analysis on Dataset B yielded similar results of Dataset A: out of 156 commented vandalisms, 124 were done by IP users consistent with the previous finding. The reasons why these pages are being vandalized are sometimes obvious, such as political reasons, religious reasons, substantial reasons, personal belief reasons, and reasons regarding immature editing on pages describing subjects such as articles pertaining to excretion, profanity, and sex, and commonly visible pages such as Wikipedia-related pages. If receives high volume of revert edits, then the page will most likely be listed at Wikipedia Proctored pages either full or semi-protection.

4.3 Community reaction (RQ3)

The community reaction is marked by the percentage of posts by anonymous users getting reverted. If $i < j < k$ in chronological order of revisions and $i = k$; then article j is a revert. To study the community reaction, 150 randomly chosen articles were sampled and the findings are depicted in Figure 2. The result of Dataset A is evident that both type of users are mainly doing normal edits during their lifetime. Registered users are making more reverts (caused by vandalism) than anonymous users and anonymous users edits are more likely to get reverted. Another important measure of community reaction is the average time elapsed before the vandalized article gets reverted. Out of these 150 sampled articles, 32 were found to be vandalized. About 25% of them were corrected in less than 90 seconds. The mean response time was about 5 minutes. The results are depicted in Figure 3. A study on Wikipedia showed about 80% of vandalism are done by unregistered users [16]. However, 81.9% of edits by unregistered users were not vandalism. It is a common misconception that all IP users are disruptive and hence their additions are routinely reverted introducing a community bias.

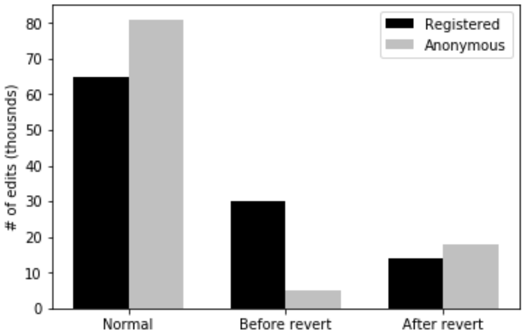

Figure 2: Community reaction between registered and anonymous users in three stages, normal is not vandalized edit, before is when a user reverting an edit, after is when an edit was reverted

5 DISCUSSION

The results of this work are fairly pre-conclusive. Registered users, as expected, account for most edits while anonymous users cause most vandalism. We observed that not all anonymous users are vandals though. Another important finding was the reliability of edits made by these users. It varied gradually over a period of time and proved that anonymous users with less number of edits are

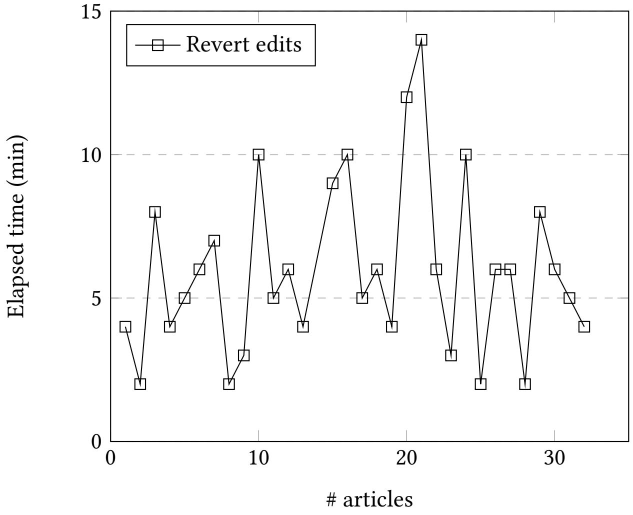

Figure 3: Elapsed time before edit is reverted

in fact more reliable than registered users. This could attribute to the fact that such unregistered users are experts on specific fields and do not bother about reputation in such community. The preliminary results showed that the community does not tolerate these misdeeds; they get reverted eventually in quick succession. Wikipedia has a counter-vandalism unit (CVU) responsible for detecting and correcting vandalism. However, the vast size of the dataset of Wikipedia requires efficient algorithm for faster detection of such anomalies. One possible solution could be predicting vandals early and keep every page semi-protected so that not every edit is reflected before scrutinized by a bot reviewer. The feasibility of such implementation requires more rigorous data mining, which is our for future work. The restrictions on dataset might introduce bias in the study. The selection of random articles was made under the assumption that the randomness of drawing the sample is purely random without any guarantee. There was no statistical test performed to test the hypotheses and approximate a confidence level.

6 CONCLUSION & FUTURE WORK

The open nature of such online discussion community has given it the utmost popularity in terms of community contribution. In this study, we aim to make Wikipedia a better place for Wikipedians by understanding how different types of users behave and how the community is reacting to such behaviour of vandalism. In this study it was evident that not necessarily all the time unregistered users mostly cause the vandalisms, and the community is particularly harsh in maintaining the content quality. This study can be a first step to solving existing issues with vandalism and ways to addressing them. This study engendered a lot of new horizon yet to be explored; few of these include:

- Why some articles are more prone to vandalism? What is the motivation behind such malice?
- Evolution of vandals over time. How is their activity throughout the day?
- Is it fruitful to block IP users from making further edits?
- Classifying the vandals to detect the vandalism ahead of time.
- Demographics of vandals and proportion of vandals using dynamic IP making them hard to catch

REFERENCES

[1] B Thomas Adler, Luca De Alfaro, Santiago M Mola-Velasco, Paolo Rosso, and Andrew G West. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 277–288.

[2] Yair Amichai-Hamburger, Naama Lamdan, Rinat Madiel, and Tsahi Hayat. 2008. Personality characteristics of Wikipedia members. *CyberPsychology & Behavior* 11, 6 (2008), 679–681.

[3] Denise Anthony, Sean W Smith, and Timothy Williamson. 2009. Reputation and reliability in collective goods: The case of the online encyclopedia Wikipedia. *Rationality and Society* 21, 3 (2009), 283–306.

[4] Harriet H Christensen, Darryll R Johnson, and Martha H Brookes. 1992. Vandalism: research, prevention, and social policy. *Gen. Tech. Rep. PNW-GTR-293. Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station. 277 p* 293 (1992).

[5] Julie E Cohen. 1995. Right to Read Anonymously: A Closer Look at Copyright Management in Cyberspace, A. *Conn. L. Rev.* 28 (1995), 981.

[6] Stefan Heindorf, Martin Potthast, Gregor Engels, and Benno Stein. 2017. Overview of the Wikidata Vandalism Detection Task at WSDM Cup 2017. *arXiv preprint arXiv:1712.05956* (2017).

[7] Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2017. Spatio-Temporal Analysis of Reverted Wikipedia Edits.. In *ICWSM*. 122–131.

[8] Michael D Lieberman and Jimmy J Lin. 2009. You Are Where You Edit: Locating Wikipedia Contributors through Edit Histories.. In *ICWSM*.

[9] Salvador Mandujano. 2003. Towards the preservation of a key feature of the internet: Policy and technology for cyberspace anonymity. In *Center for Intelligent Systems (CSI) of ITESM. In proceedings of the 7th International Conference on Technology Policy and Innovation (ICTPIâĂŽ03). Monterrey, Mexico.< http://citeseerx. ist. psu. edu/viewdoc/similar.*

[10] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 2009. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work*. ACM, 51–60.

[11] Joseph Michael Reagle. 2010. *Good faith collaboration: The culture of Wikipedia*. MIT Press.

[12] Ruqin Ren and Bei Yan. 2017. Crowd Diversity and Performance in Wikipedia: The Mediating Effects of Task Conflict and Communication. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6342–6351.

[13] Adele Santana and Donna J Wood. 2009. Transparency and social responsibility issues for Wikipedia. *Ethics and information technology* 11, 2 (2009), 133–144.

[14] Amir Sarabadani, Aaron Halfaker, and Dario Taraborelli. 2017. Building automated vandalism detection tools for Wikidata. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 1647–1654.

[15] Michael Tsikerdekis. 2013. The effects of perceived anonymity and anonymity states on conformity and groupthink in online communities: A Wikipedia study. 64 (05 2013), 1001–1015.

[16] Fernanda B Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 575–582.

[17] Adam Wierzbicki, Piotr Turek, and Radoslaw Nielek. 2010. Learning about team collaboration from Wikipedia edit history. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*. ACM, 27.