Peer-Produced Moderation: The Tradeoffs of Page Protection on Wikipedia

Leah Ajmani
University of Minnesota
Minneapolis, USA
ajman004@umn.edu

Nick Vincent
University of California, Davis
Davis, USA
nmvincent@ucdavis.edu

Stevie Chancellor
University of Minnesota
Minneapolis, USA
steviec@umn.edu

Abstract

Page protection on Wikipedia is a mechanism where the platform's core values conflict, but there is little quantitative work to ground deliberation. We introduce a case study on page protection amongst Internet Culture articles on Wikipedia and find that protected articles experience dramatic editor dropoff. These results inform future work that asks: (1) who does page protection on Wikipedia affect? and (2) is page protection effective?

Keywords: Content Moderation; Community Health; Social Media/Online Communities; Empirical study that tells us about how people use a system; Quantitative Methods

Introduction

Participation is the crux of Wikipedia's ongoing and future success; consequently, it is crucial to understand the empirical consequences of moderation techniques to ensure they have the desired effects on participation and the platform's values.

Wikipedia is an example of a platform where the consequences of friction have significant impacts on participation, such as scaring away newcomers (Halfaker et al., 2011) and minority editors (Ford and Wajcman, 2017). In fact, the Wikimedia Foundation has prioritized onboarding more diverse editors to the platform (Lam et al., 2011) as a response to this concern. Conversations about what to *do* with frictions are incomplete without exploring the *consequences* of those frictions. Understanding what happens when these frictions are deployed is crucial to facilitating constructive discussions about what they do and whether they should be used across social platforms.

In this paper, we ask: how does an exclusionary moderation technique (page protection) impact the editor landscape of an article? We show initial results from a case study that explored shifts in the editor landscape on protected pages in the Internet Culture category on Wikipedia. We propose building off of this case study by using quasi-causal methods to study the tradeoffs of protecting an article on Wikipedia.

1 Case Study: Internet Culture on Wikipedia

In Figures 1 and 2 we present the initial results of evaluating the impact of page protection on Wikipedia. Specifically, we explore the impact of page protection on protected articles within the Internet Culture category on Wikipedia (n = 108). We define **dropoff** as the number of editors who did not re-edit the page after protection and **difference** as the number of pre-protection editors minus the number of post-protection editors.

Our results suggest that protected articles experience a dramatic dropoff. Editor dropoff *increases* over time, as 75-85% of users who edit before the intervention do not return afterward (Figure 1). These results cannot be attributed to removing IP users alone; the average percentage of anonymous users on an article in our sample is 24.75%. When exploring user difference, page protection does not have a consistent positive or negative effect on total editors. Figure 2 shows the substantial spread in the underlying distribution of user difference on protected articles.

2 Proposed Future Work

From our initial results on how protected pages experience user dropoff, we propose two research questions to unpack page protection that consider the tradeoffs. For each hypothesis, we suggest specific metrics that we plan to explore through Hill & Shaw's (Hill and Shaw, 2021) quasi-causal methods.

2.1 Who does page protection on Wikipedia affect?

Our initial results suggest that page protection is affecting more than just IP editors, but where do these editors lie within the Wikipedia community? We propose two hypotheses based on our work:

(H1) Page protection brings "power users" to enter the editor landscape. As previous work suggests, power users on Wikipedia have a deep understanding of the intricate rules of Wikipedia. By the time a page is protected, in most cases, Wikipedia rules have already been violated. We plan to explore this hypothesis by using Panciera et al's (Panciera et al., 2009) definition of a power user on Wikipedia. If proven, this hypothesis would identify power users as a type of "first responder" on Wikipedia.

(H2) Page protection exacerbates barriers to the contribution that marginalized identity groups already experience, such as female editors. These barriers can be implicitly embedded in fundamental Wikipedia values, such as their five pillars (Menking and Rosenberg, 2021). We argue that page protection injects more bureaucracy and into a page, which may increase exclusion in the editor landscape. Similar to H1, we plan to explore this through a quasi-causal analysis of pages that are protected on Wikipedia on the self-disclosed gender of article editors.

2.2 Is page protection effective?

Finally, it is important to acknowledge the effectiveness of the content moderation strategy. Recall that our work is not motivated by normatively judging page protection. Rather, we seek to understand what are the tradeoffs of protecting a page to help ground future deliberation about the mechanism.

(H3) Page protection decreases vandalism. Anti-vandalism is one of the main goals of page protection. Previous work has explored how vandals are often also IP editors (Geiger and Ribes, 2010). However, does page protection accomplish this goal? Due to the massive dropoff we found in our case study, we hypothesize that page protection does meet its intended goal of thwarting vandals. We plan to explore vandalism through ClueBot activity, as it is a reliable vandalism detection bot and a common benchmark in previous work (Wang and McKeown, 2010).

(H4) Page protection increases high-quality edits. We predict that page protection increases substantive, high-quality edits on an article. After a page is protected, every editor on the article has previous edit experience and, therefore, a baseline understanding of substantive contribution. Furthermore, editors no longer have to spend time and energy policing vandals on the page. We plan to measure contribution quality through the *persistent word revisions (PWR)* as it is a widespread measure.

3 Discussion

> “Wikipedia is free content that anyone can use, edit, and distribute” -Five Pillars of Wikipedia

Page protection on Wikipedia is a moderation mechanism that has a transparent conflict with the platform’s fundamental policy: after a page is protected, only certain user groups can edit an article. If effective, page protection can serve as an example of a lightweight moderation technique that is still community-oriented. However, this raises a larger question of how content moderation affects pre-existing community barriers. Is this a necessary tradeoff to protect information quality? Are there systems that could be incorporated, such as page protection juries, that could keep the good of page protection while mitigating the bad?

References

[Ford and Wajcman2017] Heather Ford and Judy Wajcman. 2017. ’anyone can edit’, not everyone does: Wikipedia’s infrastructure and the gender gap. *Soc. Stud. Sci.*, 47(4):511–527, August.

[Geiger and Ribes2010] R Stuart Geiger and David Ribes. 2010. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW ’10, pages 117–126, New York, NY, USA, February. Association for Computing Machinery.

[Halfaker et al.2011] Aaron Halfaker, Aniket Kittur, and John Riedl. 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, pages 163–172, New York, NY, USA, October. Association for Computing Machinery.

[Hill and Shaw2021] Benjamin Mako Hill and Aaron Shaw. 2021. The hidden costs of requiring accounts: Quasi-Experimental evidence from peer production. *Communic. Res.*, 48(6):771–795, August.

[Lam et al.2011] Shyong (tony) K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP:clubhouse? an exploration of wikipedia’s gender imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym ’11, pages 1–10, New York, NY, USA, October. Association for Computing Machinery.

[Menking and Rosenberg2021] Amanda Menking and Jon Rosenberg. 2021. WP:NOT, WP:NPOV, and other stories wikipedia tells us: A feminist critique of wikipedia’s epistemology. *Sci. Technol. Human Values*, 46(3):455–479, May.

[Panciera et al.2009] 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*, pages 51–60.

[Wang and McKeown2010] William Yang Wang and Kathleen R McKeown. 2010. “got you!”: automatic vandalism detection in wikipedia with web-based shallow syntactic-semantic modeling. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING ’10, pages 1146–1154, USA, August. Association for Computational Linguistics.

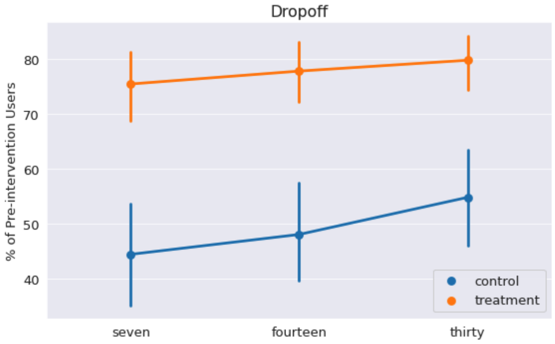

Figure 1: Normalized participant dropoff for protected articles (orange) and comparable unprotected articles (blue).

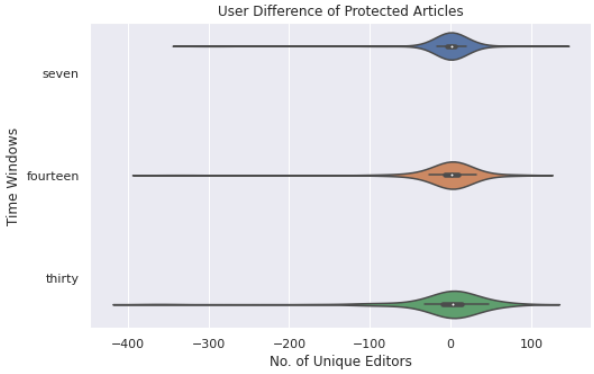

Figure 2: Kernel density estimate of user difference amongst a sample of protected articles ($n = 108$). The underlying distribution of user difference has an extreme spread [-400, 100], suggesting that it's unpredictable whether more or fewer users will edit a page after protection. Negative values signal that more users edited after page protection than before.