ICIS 2022 Proceedings

Social Media and Digital Collaboration

Dec 12th, 12:00 AM

Prominence Reduction versus Banning: An Empirical Investigation of Content Moderation Strategies in Online Platforms

Maya Mudambi
University of Maryland, mmudambi@umd.edu

Siva Viswanathan
University of Maryland College Park, sviswana@rhsmith.umd.edu

Prominence Reduction versus Banning: An Empirical Investigation of Content Moderation Strategies in Online Platforms

Short Paper

Maya Mudambi
University of Maryland, College Park
3330 Van Munching Hall
7699 Mowatt Ln, College Park MD
20742
mmudambi@umd.edu

Siva Viswanathan
University of Maryland, College Park
4313 Van Munching Hall
7699 Mowatt Ln, College Park MD
20742
sviswan1@umd.edu

Abstract

Online platforms have adopted various types of content moderation strategies to combat antisocial behaviors such as verbal aggression. This study focuses on two types of strategies: group prominence reduction and banning. This study aims to provide a holistic picture of all downstream effects of these strategies. Additionally, we assess the differential effects of content moderation on multihoming versus non-multihoming users. Preliminary findings indicate that prominence reduction strategies applied to a problematic group have the adverse effect of increasing verbal aggression in outside spaces. Banning strategies differentially impact multihoming versus non-multihoming users. These findings have important implications, as they show that group prominence reduction strategies produce negative spillover effects, and the behavior of multihoming users on multiple external platforms, and whether our results generalize across multiple contexts.

Keywords: Content moderation strategies, prominence reduction, banning, verbal aggression, multihoming

Introduction

Online platforms have adopted various types of content moderation strategies to combat antisocial behaviors such as verbal aggression, as well as misinformation. This study focuses on two types of strategies applied at the group-level: prominence reduction and banning. Prominence reduction strategies reduce the visibility and spread of harmful content, rather than deleting it. Banning strategies delete harmful content off a platform or kick problematic users out. Various studies show that both of these strategies have the potential to decrease the amount of harmful content on the focal platform (Copland 2020; Chandrasekharan et al. 2017; Chandrasekharan et al. 2022; Mudambi et al. 2022).

When evaluating content moderation strategies, it is critical to consider the entire Internet ecosystem. User-generated content (UGC) platform multihoming (simultaneously operating accounts on multiple platforms) is practiced by most Internet users (Tandoc et al. 2019). Indeed, it has been shown that content moderation strategies can have the adverse effect of pushing problematic users to external platforms (Newell et al. 2016; Copland 2020; Ali et al. 2021; Ribeiro et al. 2021). Though studies have explored the effects of both prominence reduction and banning, our study aims to provide a holistic picture of the effects of these strategies on the entire enacting platform, as well as on external platforms. We ask: how does prominence reduction versus a ban on a problematic community affect user behavior? How do these effects of group-level content moderation strategies vary for multihoming versus non multihoming users?

Preliminary findings show that when prominence reduction strategies are applied to a problematic group, they have the adverse effect of increasing verbal aggression in outside spaces, as well as the dispersion of verbal aggression on the entire platform. Furthermore, group banning strategies have different effects for multihoming versus non-multihoming users. Multihoming users decrease verbal aggression everywhere on the platform, except in topically and ideologically similar spaces. Conversely, non-multihoming users exhibit no changes in verbal aggression on the platform. Our findings have important implications for scholars and practitioners, as they provide evidence that prominence reduction strategies produce negative spillover effects elsewhere on a platform, and that the efficacy of group banning strategies may be limited.

Future work aims to elucidate any secondary spillover effects of this increase in verbal aggression in outside topically and ideologically similar forums. Furthermore, we plan on assessing how multihoming users change their behavior on multiple external platforms, to provide a more complete picture of how content moderation strategies impact user behavior on the Internet as a whole.

Background

Content Moderation Strategies

Online platforms have adopted a variety of content moderation strategies to combat the proliferation of harmful UGC (Edgecomb 2019; Benjakob 2020). This study compares the effect of two strategies applied to a problematic group: prominence reduction and banning. Prominence reduction aims to decrease the visibility of harmful UGC, rather than deleting or prohibiting it. Because problematic speech is not removed, the prominence reduction approach has the advantage of preserving free speech on a platform.

Banning strategies involve banning problematic users, groups, or topics from a platform, or deleting problematic posts. This approach has the disadvantage of raising free speech concerns, and the potential of angering users or even prompting users to leave a platform (Newell et al. 2016; Shen & Rose et al. 2019; Chandrasekharan et al. 2017). Specifically in this study, we focus on a group-level ban of users.

Study Context

Our study focuses on Reddit, a social news aggregator platform, with more than 430 million active users. The platform is composed of thousands of topic-specific forums. Each forum is governed by its own set of rules, which are enforced by volunteer user moderators. The platform as a whole is governed by site administrators who are employed by Reddit.

To assess the effects of group-level prominence reduction and banning in the presence of multihoming behavior, this study examines a single problematic Reddit forum: r/The_Donald, a forum that was dedicated to promoting Donald Trump. Due to violations of Reddit's content policy[1] regarding hate speech and threats, Reddit utilized a prominence reduction strategy in June of 2019, quarantining the forum (Robertson 2019; Vigdor & Chokshi 2019). Quarantined forums do not show up in searches, do not generate advertising revenue, and require users to see a warning and explicitly opt-in to view them. Additionally, qQuarantined forums' content will not appear in other users' newsfeed unless they were subscribed pre-quarantine. This strategy effectively isolates a problematic forum from the rest of the site.

After this initial quarantine event in June of 2019, several moderators of r/The_Donald began creating a standalone forum, entirely separate from Reddit: thedonald.win[2]. thedonald.win apes the functionality and design of the Reddit platform, but bypasses its content moderation, which was viewed by many users within this community as biased and overly restrictive (Ribeiro et al. 2020). This new forum went live on November 21, 2019. We henceforth refer to thedonald.win as the standalone forum in this paper, and Reddit's Donald Trump forum as r/The_Donald.

[1] https://www.redditinc.com/policies/content-policy

[2] The platform has subsequently changed its name to patriots.win

Due to the continuing presence of hate speech and threats, r/The_Donald was placed into restricted mode in February of 2020. Within the restricted mode, the vast majority of r/The_Donald users lost the ability to make new posts within the forum. This functioned as a group ban of the forum, as contributions effectively dropped to (and remained at) zero. The purpose of this research is to compare the effect of applying a prominence reduction strategy, versus a ban, to a problematic group on an online platform with the presence of an external alternative platform.

Related Work

In recent years there has been an explosion of work studying the effects of various content moderation strategies. Numerous papers have focused on the effect of Reddit strategies, and even the effect of Reddit's interventions upon r/The_Donald specifically. Studies show that after quarantine is applied to a problematic forum, there is a significant drop in novel users within said forum, as well as total contribution (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022). However, within forums specifically quarantined for toxic speech, there is little evidence that quarantine decreases the proportion of problematic language (Copland 2020; Chandrasekharan et al. 2022; Trujillo & Cresci 2022).

Users exposed to quarantine decrease their overall activity levels on the site (Chandrasekharan et al. 2022; Trujillo & Cresci 2022). There is conflicting evidence, however, on the effect of quarantine on censured behaviors in these outside spaces. Trujillo & Cresci (2022) found that users did not significantly change the toxicity of their posts in outside forums. However, when looking at users from communities censured for misinformation, there is evidence that users decrease misinformation spreading in outside unrelated forums, but increase it in topically and ideologically similar forums (Mudambi et al. 2022).

There have been studies examining the impact of group banning strategies upon user behavior. Banning a problematic group on a platform has been shown to significantly decrease the activity of users elsewhere on the platform (Chandrasekharan et al. 2017; Trujillo & Cresci 2022). However, there is no consensus regarding the effect of a group ban upon toxic speech elsewhere on the platform. When looking at the effect of a forum ban upon user activity on Reddit as a whole, Chandrasekharan et al. (2017) found that users significantly decreased toxic speech usage. Conversely, Trujillo & Cresci (2022) found that r/The_Donald users did not significantly change the median toxicity of their posts on the platform.

Researchers have established a link between content moderation (through either banning or prominence reduction) and user migration to external platforms with less stringent policies (Newell et al. 2016; Copland 2020; Ali et al. 2021; Ribeiro et al. 2021). There is limited evidence for pure migration behavior; instead, most users multihome, maintaining activity on the focal platform, and the new external platform (Newell et al. 2016). Multihomers have been shown to increase toxicity on external platforms (Ribeiro et al. 2020).

Research Gap

In this work we aim to contribute an understanding of how prominence reduction and banning strategies differ in their effect upon users, all while considering the effect of multihoming. Though there is evidence that prominence reduction curtails engagement within directly impacted communities (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022), there are conflicting findings regarding its effect upon user behavior elsewhere on a platform. Trujillo & Cresci (2022) examined the behavior of r/The_Donald users post-quarantine and found no change in toxicity elsewhere on Reddit, but did not partition this analysis by the type of forum. This is significant, because previous research shows that after prominence reduction, users factor in topic and ideology when moving to new spaces on a platform (Mudambi et al 2022). Group banning has been shown to decrease the user activity (Chandrasekharan et al. 2017; Trujillo & Cresci 2022). However, there are mixed results as to the effects of a ban upon verbal aggression.

Content moderation has been shown to push users onto external platforms with laxer policies (Newell et al. 2016; Copland 2020; Ali et al. 2021; Ribeiro et al. 2021). Most of these users multihome and concurrently operate accounts on the old and new platforms (Newell et al. 2016). Previous research has examined the effect of a group ban upon these users (Ribeiro et al. 2020), but it only described user trends within the external platform, not the platform enacting content moderation. We aim to ultimately provide a fuller understanding of how multihoming users respond to content moderation across multiple platforms.

This work makes theoretical contributions to the literature on community-level content moderation, online verbal aggression, and online multihoming. We address key research gaps regarding the downstream effects of group prominence reduction as compared to group banning. We also address research gaps on online verbal aggression. Finally, by factoring in multihoming behavior, we contextualize online content moderation strategies within the entire Internet ecosystem. This addresses a key research gap, as most previous work examines the impact of content moderation strategies on the enacting platform only.

This research addresses methodological research gaps. There are studies examining the effect of prominence reduction and banning, but most describe behavioral trends within the directly impacted group only, without comparison to a control group (Copland 2020; Ribeiro et al. 2020; Trujllo & Cresci 2022). To contrast, we collect data for a control group, and utilize a difference-in-difference specification. Previous studies assess changes in general verbal toxicity (Ali et al. 2021; Ribeiro et a. 2021; Trujillo & Cresci 2021), which captures behaviors from profanity to hate speech. In this work we focus on behavior specifically censured by Reddit (hate speech and threats) to better assess users' responses to platform intervention.

Theory & Hypotheses Development

We draw from network effects formulate our initial hypothesis. Contribution volume on online platforms is strongly influenced by network effects. This means that the value that an individual gains is directly proportional to the number of other users present (Zhang et al. 2012). Studies show that group size, as well as the number of social ties available in a social network is tied to the frequency of user contribution (Burke et al. 2009; Zhang & Zhu 2011; Shriver et al. 2013; Baek & Shore 2020). Given that quarantine drastically reduces the visibility of a forum, we predict a drop in both group-size and individual contribution within the affected forum. This is in-line with previous research on group-level prominence reduction strategies (Copland 2020; Chandrasekharan et al. 2022; Mudambi et al. 2022). Thus:

> **Hypothesis 1**. After group prominence reduction, affected users will decrease verbal aggression within the group.

Following this reduction in verbal aggression within r/The_Donald, we predict users will seek out other audiences on the platform to replace this behavior. We draw from self-presentation and social norms theory to formulate our hypotheses regarding behavior elsewhere on the platform. The foundational tenet of self-presentation theory is that individuals tailor their social performances to their audience (Goffman 1959). For users of an online platform, however, the actual audience is unclear. Thus, the concept of the imagined audience – or a person's conception of the people witnessing their behavior – more directly applies. The behavior of users on online platforms is directly influenced by the characteristics of the imagined audience with whom they are communicating (Litt 2012).

The composition of a user's imagined audience on an online platform influences their subsequent behavior (Litt 2012). It stands to reason that the audience present in outside right-wing forums will be imagined by r/The_Donald users to be the most similar to that in the quarantined forum, and user behavior in these forums post-quarantine will be influenced by this conception. Previous research shows that quarantine-exposed users increased platform-censured behavior in outside forums ideologically closest to the focal quarantined forum (Mudambi et al. 2022).

Additionally, at the time of the quarantine of r/The_Donald, Reddit had already established a social norm in their content policy: the prohibition of hate speech, as well as violent threats. The widely publicized quarantine of that forum explicitly for the violation of this norm made it more salient to the entire user base. Given online users' tendency to conform to platform-wide toxicity norms (Kwon & Gruzd 2017; Rajadesingan et al. 2020), we predict:

> **Hypothesis 2**. After prominence reduction, affected users will only increase verbal aggression in outside spaces of a similar ideology.

Group bans have been shown to decrease the overall activity of users on a platform (Trujillo & Cresci 2022), but can also push users into increasing activity in select spaces (Chandrasekharan et al. 2017). After the restriction of r/The_Donald, the social norm against hate speech and threats was again stressed.

Post group ban, multihomers have been shown to increase toxicity on external platforms (Ribeiro et al. 2020). Indeed, the general level of toxicity of these external platforms has been shown to be much higher

(Ribeiro et al. 2020; Li et al. 2021). Given the higher toxicity norm on the standalone forum (thedonald.win), and the high degree of similarity between the topic and ideology between the standalone forum and r/The_Donald, we expect that multihoming users will increase verbal aggression on the standalone forum, rather than Reddit right-wing forums. Thus:

> **Hypothesis 3**. After a group ban, only non-multihoming users will increase verbal aggression in outside forums of a similar ideology.

Data

Data Collection

Reddit quarantined r/The_Donald on 6/26/19 and placed it in restricted mode on 2/26/20. For the quarantine analysis, our treatment group is users who posted in r/The_Donald in a baseline period 3 months prior to quarantine. Our control group posted in right-wing forums (r/Conservative and r/Libertarian)[3] in this period. We scraped all posts for these users +/- 3 months around quarantine. We eliminated control users that had ever posted in a quarantined or banned forum, as well as any bots.

To obtain preliminary results regarding the ban, we completed the same protocol, but for a random pilot sample of treatment and control users. For all treatment users in this sample, we assessed whether they created an account on the standalone forum, thedonald.win. As in previous research on multihoming, users were matched assuming they retained their Reddit username (Newell et al. 2016; Ribeiro et al. 2020; Ali et al. 2021). This is likely, as thedonald.win had a feature that allowed users to reserve their Reddit username to maintain continuity between the platforms (Ribeiro et al. 2020). Users who made an account on the standalone forum are multihomers, and users with no evidence of standalone forum account are non-multihomers.

Identifying Verbal Aggression

The verbal aggression types of interest in this study are hate speech and threats. They were the specific behaviors cited by Reddit administrators when placing r/The_Donald in quarantine, as well as its later restriction (Robertson 2019; Vigdor & Chokshi 2019).

We utilized the Google Perspective API[4] to identify hate speech and threats within the dataset. This API utilizes machine learning models trained on corpora of human-annotated text and can produce probability scores for four distinct types of verbal toxicity (insults, profanity, hate speech, threats), and two general types of toxicity (toxicity, severe toxicity,). There is a high correlation between the probability scores produced by Google Perspective for the six types of verbal aggression[5]. In contrast with previous studies, which focus on general toxicity scores produced by the Perspective API (Ali et al. 2021; Ribeiro et a. 2021; Trujillo & Cresci 2021), we focus on the behavior specifically censured by Reddit: hate speech and threats.

For a post to be classified as hate speech in this study, its probability score for hate speech had to be higher than the three other distinct types of verbal aggression, as well as above 75%. The same was true for threats. Any post classified as either hate speech or a threat is deemed a verbal aggression post (VA).

Empirical Analysis

We utilize difference-in-difference (DID) analysis to analyze the effect of prominence reduction, as well as a group ban, upon user behavior. In this DID analysis, we compare the behaviors of our sample of r/The_Donald users (treatment group) with a control group of other Reddit users. We have two interventions of interest: exposure to quarantine, and a group ban, and thus run two types of analyses. In

[3] These communities were selected due to a similarity with r/The_Donald in terms of political ideology, as well as the usage of verbal aggression. Users in both treatment (3.6%) and control (3.0%) groups posted a similar proportion of verbal aggression posts on the platform during the baseline period prior to quarantine.

[4] https://www.perspectiveapi.com/

[5] Correlations on verbal aggression probability scores were run on a sample of 199,014 user posts from the quarantine baseline period (3/26/19 – 6/26/19).

both cases, the treatment group was exposed to the intervention, and the control group was not. For both analyses, our data comprises user-level posting data +/- 3 months before and after the intervention of interest (quarantine or group ban), as well as user account features.

We ran DID OLS regressions with robust standard errors on a weekly panel to assess the relative effects of quarantine upon verbal aggression posting (VA). For this specification, i represents the user, and j represents the week. The dependent variable *VA*$_{ij}$ represents verbal aggression posts. The dummy variable *TheDonald*$_i$ indicates whether user *i* is in the treated group. The variable *PostQuarantine*$_j$ indicates whether the period is after the quarantine. The interaction term *TheDonald*$_i$ *X PostQuarantine*$_j$ indicates the effect of the intervention on the behavior of the treated users. The variables *Moderator*$_i$*, PremiumMember*$_i$*, VerifiedEmail*$_i$*, Tenure*$_i$ and weekly dummies, are the control variables.

$$VA_{ij} = (TheDonald_i \; X \; PostQuarantine_j) + TheDonaldi_i + PostQuarantine_j + Moderator_i + PremiumMember_i + VerifiedEmail_i + Tenure_i + \sum_{j=0}^{J} Week_j + \varepsilon_{ij}$$

For our preliminary ban analyses, we used a similar specification, except only comparing baseline and follow-up. In this case, *j* represents the period, rather than the week.

Preliminary Findings

The Effect of Quarantine on Verbal Aggression

Descriptive analysis shows that on average, users decrease verbal aggression posts from 0.049 to 0.037 after quarantine within r/The_Donald, a drop of 23.2%. **H1** is supported.

We then model the number of verbal aggression posts (VA) in outside right-wing, centrist, and non-political forums on Reddit. The coefficient estimates are shown below in Table 1; robust standard errors are in parentheses and p-values are below. We find evidence that r/The_Donald users exposed to quarantine increase verbal aggression posts within outside right-wing forums. The coefficient on the interaction term, *TheDonald*$_i$ *X Post-Quarantine*$_j$, is positive (0.006) and significant (p<0.000), showing that users exposed to quarantine increase verbal aggression in topically and ideologically similar right-wing forums. However, contrary to expectation, we find that these users also increase verbal aggression in outside centrist (β=0.003; p<0.000) and non-political forums (β=0.073; p<0.000), as well. **H2** is partially supported.

Given that treated users increase VA in a variety of forums after quarantine, we ran a supplementary analysis to examine any changes in misinformation dispersion (as measured by the Herfindahl-Hirschman index) on the entire platform. The coefficient on the interaction term, *TheDonald*$_i$ *X Post-Quarantine*$_j$, is negative (-174.7) and significant (p<0.000), showing that users increase their VA dispersion.

	VA Posts			**VA Dispersion**
Location	**Centrist Forums**	**Outside Right Forums**	**Non-Political Forums**	**Reddit**
Quarantine-Exposed$_i$ X Post-Quarantine$_j$	0.003 (0.001) 0.000	0.006 (0.001) 0.000	0.073 (0.003) 0.000	-174.7 (22.35) 0.000
Quarantine-Exposed$_i$	-0.0134 (0.001) 0.000	-0.0190 (0.001) 0.000	-0.260 (0.002) 0.000	1,215 (15.31) 0.000
Post-Quarantine$_j$	-0.007 (0.001) 0.000	-0.008 (0.001) 0.000	-0.114 (0.003) 0.000	363.1 (49.20) 0.000
Is Moderator$_i$	0.001 (0.000) 0.025	0.002 (0.000) 0.000	0.048 (0.001) 0.000	-265.2 (14.65) 0.000
Premium Account$_i$	0.004 (0.001) 0.003	0.006 (0.001) 0.000	0.052 (0.006) 0.000	-319.2 (59.60) 0.000
Verified Email$_i$	-0.002 (0.000) 0.000	-0.002 (0.000) 0.211	0.002 (0.001) 0.026	159.1 (14.04) 0.000
Tenure$_i$	0.000 (0.000) 0.000	0.000 (0.000) 0.772	0.000 (0.000) 0.000	0.158 (0.006) 0.000

+ ***week dummies***				
Constant$_{ij}$	0.016 (0.001) 0.000	0.019 (0.001) 0.000	0.296 (0.003) 0.000	8,130 (33.36) 0.000
Observations	2,475,468	2,475,468	2,475,468	142,470
R^2	0.003	0.004	0.036	0.072
Robust standard errors in parentheses; P-values below				

Table 1. The Effect of Quarantine upon Verbal Aggression

The Effect of Group Ban on Verbal Aggression

This analysis was performed on a random pilot sample of treatment and control users. We model the number of verbal aggression posts (VA) in outside right-wing forums on Reddit, as well as all other outside forums on the platform for non-multihomer r/The_Donald users compared to the control group. The coefficient estimates are shown below in Table 2; robust standard errors are in parentheses and p-values are below. Group ban has no discernable effect on the volume of verbal aggression posts within outside right-wing forums, or elsewhere on the Reddit platform for non-multihomer users. **H3** is not supported.

We then compare multihomer r/The_Donald users with the control group. The coefficient estimates are shown below in Table 2. We find that group ban has no discernable effect on the volume of verbal aggression posts within outside right-wing forums for multihomer users. However, multihomer users significantly ($p<0.056$) decrease verbal aggression elsewhere on the Reddit platform, which we did not predict.

	Non-Multihomers		**Multihomers**	
VA:	**Outside Right-Wing Forums**	**Elsewhere on Reddit**	**Outside Right-Wing Forums**	**Elsewhere on Reddit**
TheDonald$_i$ X PostBan$_j$	0.0543 (0.117) 0.643	-0.699 (0.567) 0.218	0.0998 (0.169) 0.554	-3.016 (1.579) 0.0561
TheDonald$_i$	-0.576 (0.0928) 5.97e-10	-0.376 (0.447) 0.400	-0.661 (0.128) 2.54e-07	2.267 (1.365) 0.0969
PostBan$_j$	-0.0998 (0.106) 0.345	-0.329 (0.320) 0.305	-0.0998 (0.106) 0.345	-0.329 (0.320) 0.304
Moderator$_i$	0.257 (0.130) 0.0487	2.770 (0.500) 3.21e-08	0.259 (0.150) 0.0849	2.039 (0.512) 6.89e-05
Premium Member$_i$	0.527 (0.545) 0.334	9.217 (2.805) 0.00102	0.934 (0.681) 0.170	12.87 (3.437) 0.000183
Verified Email$_i$	-0.0735 (0.0943) 0.436	0.192 (0.375) 0.609	-0.0991 (0.133) 0.455	-0.335 (0.501) 0.504
Tenure$_i$	0.0002 (6.8e-05) 0.000879	-0.000 (0.000) 0.162	0.0003 (8.36e-05) 0.000456	-0.000 (0.000) 0.0702
Constant$_{ij}$	0.413 (0.141) 0.00345	3.491 (0.412) 0	0.356 (0.178) 0.0459	4.077 (0.507) 0
Observations	5,518	5,518	4,342	4,342
R^2	0.014	0.027	0.12	0.031

Table 2. The Effect of Group Ban on Outside Verbal Aggression

Summary & Discussion

This study uses data from users of r/The_Donald to compare the effects of group-level prominence reduction with group banning strategies on verbal aggression. The prominence reduction strategy examined in this study is Reddit's quarantine policy; the banning strategy is Reddit restricted mode.

Consistent with previous work in network effects, we find that verbal aggression significantly decreases within r/The_Donald after quarantine. We also find evidence of a replacement effect: these users increase verbal aggression in outside right-wing, centrist, and non-political forums after quarantine. This results in an increase in verbal aggression dispersion on the platform. This is a novel finding; previous studies examining the effect of group prominence reduction on verbal aggression did not uncover this negative externality (Chandrasekharan et al. 2022; Trujillo & Cresci 2022). This suggests that users do not factor in a space's ideology, or platform-wide toxicity norms in their behavior. We conclude that group prominence reduction is ineffective in mitigating the harm caused by verbal aggression on an online platform.

We find that non-multihomers do not significantly change verbal aggression in any forums on Reddit. This suggests that repeated interventions upon the same group of problematic non-multihoming users may be ineffective. Multihoming users were not found to significantly change verbal aggression in right-wing forums after the group ban. However, they were found to significantly decrease verbal aggression elsewhere on the platform. Previous research shows that users who create accounts on external platforms increase their toxicity there (Ribeiro et al. 2020; Ali et al. 2021). This suggests that these platforms have a higher level of toxicity as the social norm. According to self-presentation theory, individuals tailor their behavior based on the audience present (Goffman 1959). Previous research shows that online users adapt their toxicity levels based on their audience, as well as social norms (Kwon & Gruzd 2017; Rajadesingan et al. 2020). Our finding suggests that this increase in toxicity on external platform(s) is accompanied with a corresponding decrease in the home platform (except in ideologically and topically similar forums).

Conclusions & Future Work

We hope to contribute a holistic view of the effects of two different group content moderation strategies upon a problematic community. Thus far, we have analyzed how group-level prominence reduction, as well as a ban, affect user behavior on a platform, and how multihoming users' behavior differs from non-multihoming users. We have multiple plans for improving and supplementing these preliminary results.

We ran our group ban analyses on a sub-sample of control and treatment users. To eliminate the possibility of sampling bias, we will gather data for all appropriate users, and run difference-in-differences analysis on a weekly panel to assess behavior changes in right-wing, centrist, and non-political outside forums.

To complete the full picture of how these content moderation strategies affect the entire Internet ecosystem, we will collect all posts on the standalone forum, thedonald.win. Additionally, there are multiple platforms such as Gab, Metafilter, and Raddle, which have been promoted on Reddit as alternatives[6]. We will attempt to match users within our sample on these alternative platforms, gather the appropriate data, and run analyses that assess how multihomer behavior on these alternatives compares with the behavior on Reddit.

Finally, it is possible that the political context of this work is unique. Future work will assess whether our findings generalize to different types of communities undergoing group content moderation strategies.

References

Ali, S., Saeed, M.H., Aldreabi, E., Blackburn, J., De Cristofaro, El, Zannettou, S., and Gianluca, S. 2021. "Understanding the Effect of Deplatforming on Social Networks", *Proceedings of the 13th ACM Web Science Conference*.

Benjakob, O. 2020. "Why Wikipedia Is Much More Effective Than Facebook at Fighting Fake News," *Haaretz*. https://www.haaretz.com/us-news/.premium-why-wikipedia-is-much-more-effective-than-facebook-at-fighting-fake-news-1.8378622.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. 2017. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," *Proceedings of the ACM on Human-Computer Interaction* (2:31).

Chandrasekharan, E., Jhaver, S., Bruckman, A., and Gilbert, E. 2022. "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit," *ACM Transactions on Computer-Human Interaction* (29:4), 29.

Copland, S. 2020. "Reddit quarantined: Can changing platform affordances reduce hateful material online?" *Internet Policy Review* (9:4), pp. 1-26.

[6] https://www.reddit.com/r/RedditAlternatives/comments/oioeot/list_of_active_reddit_alternatives_v7

Edgecomb, C. 2019. “Facebook and Instagram starting to identify and label ‘fake news’,” *Impact BND,* https://www.impactbnd.com/blog/facebook-and-instagram-starting-to-identify-and-label-fake-news-before-it-goes-viral.

Goffman, E. 1959. *The presentation of self in everyday life,* New York: Doubleday Anchor.

Kwon, K.H., and Gruzd, A. 2017. “Is Aggression Contagious Online? A Case of Swearing on Donald Trump’s Campaign Videos on Youtube,” *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Mudambi, M., Clark, J., and Rhue, L. 2022. “Fighting Misinformation on Social Media with Prominence Reduction Strategies,” *Working Paper*.

Newell, E., Jurgens, D., Saleem, H.M., Vala, H., Sassine, J., Armstrong, C., Ruths, D. 2016. “User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest,” *Proceedings of the Tenth International AAAI Conference on Web and Social M*edia, , pp. 279-289.

Rajadesingan, A., Resnick, P., and Budak, C. 2020. “Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits,” *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*.

Ribeiro, M.H., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., and West, R. 2020. “Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels”, *arXiv:2010.10397v2*.

Robertson, A. 2019. “Reddit quarantines Trump subreddit r/The_Donald for violent comments,” *The Verge*, https://www.theverge.com/2019/6/26/18759967/reddit-quarantines-the-donald-trump-subreddit-misbehavior-violence-police-oregon.

Shen, Q., and Rose, C.P. 2019. “The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit’s Quarantine Policy,” *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, pp. 58-69.

Song, Y., and Wu, Y. 2018. “Tracking the viral spread of incivility on social networking sites: The case of cursing in online discussions of Hong Kong-Mainland China conflict,” *Communication and the Public* (31), pp. 46-61.

Tandoc Jr., E.C., Lou, C., and Min, V.L.H. 2019. “Platform-swinging in a poly-social-media context: How and why users navigate multiple social media platforms,” *Journal of Computer-Mediated Communication* (24), 21-35.

Trujillo, A., and Cresci, S. 2022. “Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald,” *arXiv: 2201.06455*.

Vigdor, N., and Chokshi, N. 2019. “Reddit Restricts Pro-Trump Forum Because of Threats”, *The New York Times,* https://www.nytimes.com/2019/06/26/us/politics/reddit-donald-trump-quarantined.html.

Young, H.P. 2008. “Social norms,” in *The New Palgrave Dictionary of Economics (2nd ed.)*, London: Macmillan.