9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning

Keywords: deplatforming, online communities, Reddit, observational studies, toxicity

Extended Abstract

Motivation and Findings. Online communities have existed since the early days of the internet, where individuals could interact with each other around shared interests through bulletin boards and chat systems. However, today, they are commonly hosted on mainstream social media platforms such as Reddit and Facebook. With the increasing demand to keep these communities civil and respectful, these platforms have started to ban communities that violate their guidelines. Unfortunately, such bans can have unintended consequences as users may migrate to alternative, fringe platforms with lower moderation standards and be exposed to antisocial behaviors (Zuckerman and Rajendra-Nicolucci, 2021; Trujillo and Cresci, 2022). This spillover of behavior from the fringe platforms can then limit the effectiveness of community bans on mainstream platforms. To test whether such spillover exists, we investigate if co-active users—active on both fringe platforms and mainstream social media—become more toxic on the mainstream platform after joining a fringe platform. We find that co-active users exhibit consistent and increased antisocial behavior on Reddit.

Data. To conduct this investigation, we analyzed data from 70,000 users from the three subreddits r/The_Donald, r/Incels, and r/GenderCritical. In all three cases, after banning, users migrated en masse to alternative, fringe platforms (thedonald.win, incels.co, and ovarit.com). Thus, we collect the entire posting history for the users active in these communities (i) on Reddit and (ii) on the relative fringe platform. Overall, we collect four million posts from the three focal subreddits. Additionally, we implement custom web crawlers to collect data from *thedonald.win*, *incels.co*, and *ovarit.com*. We collect over three million posts by $42,340$ users from *thedonald.win*, *ovarit.com*, and *incels.co* in total. Finally, to understand the effect of co-participation on fringe platforms on users' behavior on Reddit, we define *co-active* users as those posting both on Reddit and the fringe platforms after the banning. We obtain $1,478$ Reddit *co-active* users on Reddit and the fringe platforms. We label all users posting on Reddit after the ban without a matching username on the fringe platform as *Reddit-only* users. We find $14,810$ *Reddit-only* users that were previously members of the three subreddits.

Methods. We compare the behavior of Reddit users who co-participate in a fringe platform with those who only use Reddit. To measure antisocial behavior, we look at users' toxicity and controversial group engagement in other extreme subreddits. To estimate causal effects, we combine two quasi-experimental causal inference methods: propensity score matching and difference-in-differences. Using one-to-one propensity score matching, we match similar co-active and Reddit-only users from the pre-banning period to mitigate the risk of differences in post-banning behavior due to user characteristics. We train a logistic regression classifier

on pre-banning user features to estimate the propensity score, and match users using the nearest neighbor algorithm. Using the matched sample from eighteen weeks before and after the subreddit ban, we use a DiD model to estimate the effect of co-activity on users' antisocial behavior on Reddit.

Results. Our study reveals that users who participate both on Reddit and fringe platforms tend to exhibit more antisocial behavior when faced with a community ban. In our analysis, we used the DiD regression to quantify the impact of co-participation in fringe platforms on Reddit antisocial behavior, as shown in Figure 1 (top-row). Specifically, we measured the difference in toxicity or engagement between users who participate on both platforms and those who only use Reddit, net of any pre-ban differences (DiD effect). Our findings indicate that the antisocial behavior of users who co-participate on both platforms diverges over time from those who only use Reddit, as illustrated in Figure 1 (top-row). Furthermore, the DiD effect becomes more pronounced with time, suggesting that co-participating users on Reddit become increasingly toxic and engage more with controversial subreddits, as demonstrated in Figure 1 (top-row). This result provides evidence that the adoption of antisocial behaviors by *co-active* users not only increases but *diverges* from that of *Reddit-only* users. To estimate the rate of divergence, we consider the *time* as an integer variable taking values in $[-18, +18]$ and not as a variable indicating pre- or post-banning period. Figure 1 (bottom row) shows the models fitted on r/The_Donald, r/GenderCritical , and r/Incels. We observe that *co-active* users diverge consistently from *Reddit-only* users in toxicity. In particular, we find that the increase in toxicity for *co-active* users exceeds that of *Reddit-only* users by 2% *weekly*. These results remain consistent across all three subreddits and for controversial group engagement. This analysis provides statistical evidence that the antisocial behavior of co-participating users not only sharply increases immediately after the ban but keeps growing at a higher rate than Reddit-Only users.

Discussion. Our results shed light on the relationship between fringe and mainstream social media. Although mainstream social media stakeholders may believe that the departure of users who exhibit antisocial behavior is beneficial, our research suggests that co-active users serve as a conduit for antisocial behavior to spill back into mainstream social media from fringe platforms. Indeed, our results indicate that users exposed to toxic environments on fringe platforms will act similarly on the mainstream platform. Our results have two critical implications for platform stakeholders. First, they suggest that antisocial behavior spills from fringe onto mainstream platforms, limiting the efficacy of such community-level bans. Second, our results provide a clear target to reduce unintended within-platform consequences of community-level bans: *co-active users*. Platforms could develop more sophisticated interventions that remove problematic communities *and* discourage co-activity. For example, when banning communities like those studied, Reddit could *also* apply sanctions to their users, such as reducing the visibility of their posts.

References

Trujillo, A.; Cresci, S. (2022). Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *CSCW22* .

Zuckerman, E.; Rajendra-Nicolucci, C. (2021). Deplatforming Our Way to the Alt-Tech Ecosystem. *Knight First Amendment Institute at Columbia University, January* **11**.

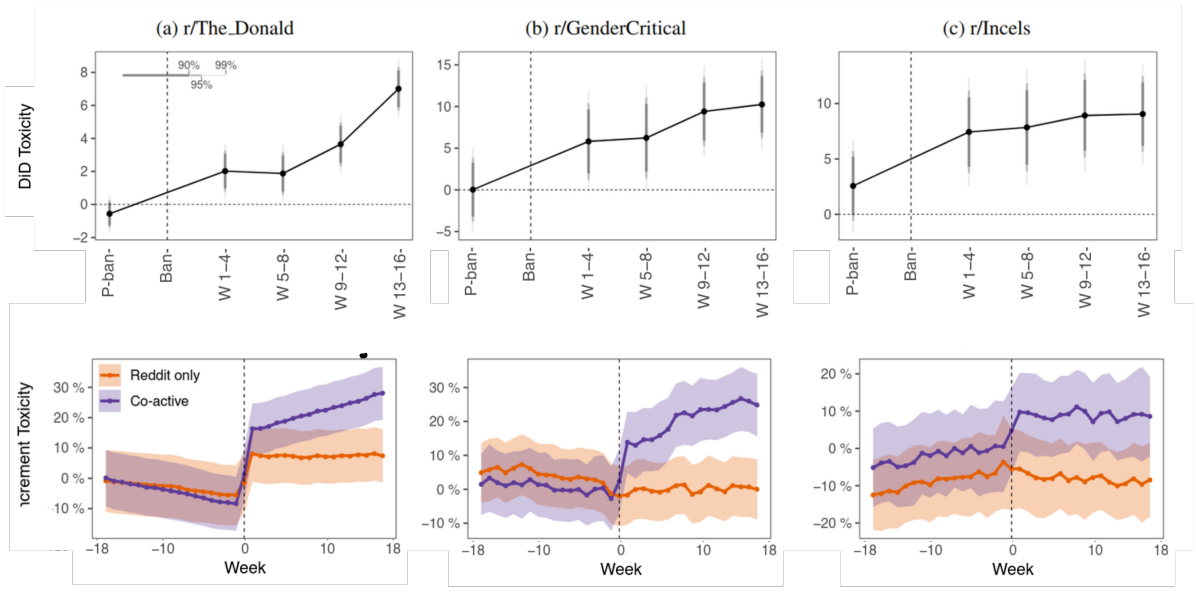

Figure 1: Estimated DiD effect of co-participation for toxicity (top row) shown for r/The Donald (fig. 1a), r/GenderCritical (fig. 1b), and r/Incels (fig. 1c). Divergence of toxicity (bottom row) for co-active (purple) and Reddit-only users (orange) shown for r/The Donald (fig. 2a), r/GenderCritical (fig. 2b), and r/Incels (fig. 2c).