

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

Giuseppe Russo^{*(1)}, Luca Verginer^{†(1)}, Manoel Horta Ribeiro^{‡(2)}, Giona Casiraghi^{§(1)}

(1) ETH Zürich, Chair of Systems Design, Weinbergstrasse 56/58, Zürich, Switzerland

(2) EPF Lausanne, DLab, Station 14, Lausanne, Switzerland

Abstract

Online platforms face pressure to keep their communities civil and respectful. Thus, the bannings of problematic online communities from mainstream platforms like Reddit and Facebook are often met with enthusiastic public reactions. However, this policy can lead users to migrate to alternative fringe platforms with lower moderation standards and where antisocial behaviors like trolling and harassment are widely accepted. As users of these communities often remain *co-active* across mainstream and fringe platforms, antisocial behaviors may spill over onto the mainstream platform. We study this possible spillover by analyzing around 70,000 users from three banned communities that migrated to fringe platforms: r/The_Donald, r/GenderCritical, and r/Incels. Using a difference-in-differences design, we contrast *co-active* users with matched counterparts to estimate the causal effect of fringe platform participation on users' antisocial behavior on Reddit. Our results show that participating in the fringe communities increases users' toxicity on Reddit (as measured by Perspective API) and involvement with subreddits similar to the banned community—which often also breach platform norms. The effect intensifies with time and exposure to the fringe platform. In short, we find evidence for a spillover of antisocial behavior from fringe platforms onto Reddit via co-participation.

1 Introduction

Online communities, “aggregations of individuals who interact around a shared interest” [31], date back to the bulletin boards and chat systems of the early days of the Web [32]. Today, thriving online communities are often hosted on mainstream social media platforms like Reddit and Facebook. Mainstream platforms moderate communities through a two-tiered governance system. The platform is responsible for coarse-grained measures, like creating guidelines that all communities should adhere to and sanctioning communities that fail to conform to them [20]. On the community level, volunteer moderators make fine-grained moderation decisions, such as determining rules specific to the community and removing posts deemed inappropriate [44].

Recently, online platforms have often banned—entirely deactivated—communities that breached their increasingly comprehensive guidelines. In 2020 alone, Reddit banned around 2,000 subreddits (the name a community receives on the platform) associated with hate speech [35]. Similarly, Facebook banned 1,500 pages and groups related to the QAnon conspiracy theory [7]. While these decisions are met with enthusiasm [e.g., see Anti-Defamation League [2]], the efficacy of “deplatforming” these online communities has been questioned [57]. When

*russog@ethz.ch, corresponding author

†lverginer@ethz.ch

‡manoe.lhortaribeiro@epfl.ch

§gcasiraghi@ethz.ch

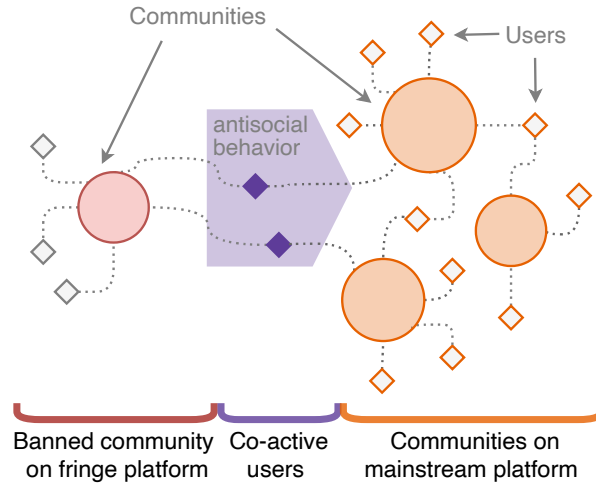


Figure 1: Motivation. When communities are banned from a mainstream platform and relocate to a fringe platform, antisocial behaviors may spill over onto the mainstream one through *co-active* users, i.e., users active across platforms. In this paper, we study this spillover effect by analyzing *co-active* users in three fringe communities banned from Reddit.

mainstream platforms ban entire communities for their offensive rhetoric, users often migrate to alternative *fringe platforms*, sometimes created exclusively to host the banned community [8]. Banning, in that context, would not only strengthen the infrastructure hosting these fringe platforms [57] but allow these communities to become more toxic elsewhere [16].

Banning online communities may also impact the mainstream platforms themselves [47]. In Fig. 1, we illustrate one mechanism by which this may happen. When problematic communities are banned, users may choose to remain active in both mainstream and fringe platforms, creating feedback between online spaces with little to no moderation and social networks. In the fringe platform, these *co-active* users are likely exposed to increased toxicity and misinformation and may participate in harassment, doxing, and defamation campaigns [12]. Consequently, antisocial behaviors from fringe platforms may spill over into other unbanned communities within mainstream social media where co-active users participate.

Present work. In this paper, we conduct a large-scale longitudinal study comparing the behavior on Reddit of users from a banned community posting also on a fringe platform to users posting only on Reddit. We find that users who co-participate—active on both platforms—exhibit more antisocial behaviors on Reddit than users posting on Reddit only. This effect intensifies over time and increases with exposure to the fringe platform. In short, we find evidence of spillovers of antisocial behavior from fringe platforms onto mainstream social media through *co-active* users.

2 Related Work

Measuring antisocial behavior on the Web. Antisocial behavior has existed on the Web since its early days [9], with users engaging in different types of behavior like *trolling*, i.e., intentionally disrupting a discussion or community [6], and *harassment*, attempts to demean or humiliate [29]. Previous works have attempted to measure the prevalence of antisocial behaviors [6, 51], as well as to understand factors that would lead users to engage in them [5].

One widely used machine learning tool to measure online antisocial behavior is Perspective API from Jigsaw [19]. It provides “toxicity” scores to posts indicating if they would lead to someone leaving a discussion due to their rude and disrespectful nature. Perspective and other automated content moderation tools have faced widespread criticism: they lack context, fail to distinguish between legitimate and rule-breaking content, and are biased against minorities [38, 42]. At the same time, Perspective has proven to be a valuable tool for researchers to study online antisocial behavior. Previous research on Reddit and Facebook data [22, 34] shows that its performance is similar to that of a human annotator. It further outperforms keyword-based alternatives [54].

Online antisocial communities. Antisocial communities are groups of users consistently engaging in antisocial behavior [25]. They are often sympathetic to conspiracy theories [e.g., QAnon [43]] and extremist ideologies [e.g., the Alt-right [37]]. They have been shown to have disproportionate influence over memes and news shared on the web [52, 53]. Further, they have been closely associated with medical misinformation, conspiracy theories, and extremist ideologies that significantly impact the real world [26, 45, 55].

Among these communities, the most relevant for this work are the following: r/The_Donald, r/GenderCritical, and r/Incel. The subreddit r/The_Donald was created in June 2015 to support the then-presidential candidate Donald Trump’s bid for the U.S. Presidential election. This community has been closely linked with the rise of the “alt-right” movement, and was known to host racist, sexist and islamophobic discussions [24] and to spread conspiracy theories [28]. Flores-Saviaga *et al.* [11] have studied how active participants in r/The_Donald mobilized the community to engage in “political trolling”. The subreddit r/GenderCritical was created in September 2013 to host the trans-exclusionary radical feminist (TERF) community. TERFs hold the view that gender derives from biological sex [50], and the community at large has consistently used social media to dox and harass trans women [21]. The subreddit r/Incel was created in August 2013 to host a community of self-denominated “*involuntary celibates*.” Incels abide by “The Black Pill,” the belief that unattractive men would be doomed to romantic loneliness and unhappiness. Previous work has studied the community links with other masculinist communities [36], as well as its relationship with terrorist attacks [15] and the production of misogynistic content online [17].

Analyzing the effects of deplatforming. Although different, a commonality between r/Incel, r/The_Donald, and r/GenderCritical is that they have been “de-platformed,” i.e., banned from Reddit for breaching their guidelines. Previous works have studied the effects of deplatforming of communities and users, finding that, following the ban, users reduce their activity on mainstream platforms [18], but also that users often migrate to other fringe platforms, where they at times become more toxic than before [1, 16]. Moreover, Trujillo and Cresci [47] have

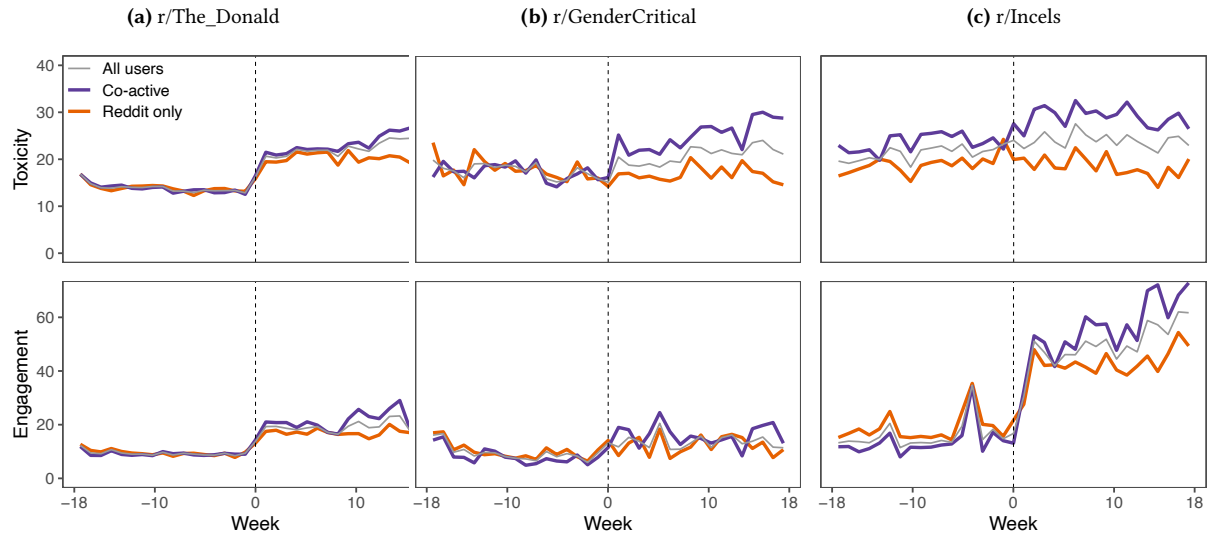


Figure 2: Toxicity mean values (top row) and Engagement mean values (bottom row) for *co-active*, *Reddit-only*, and *all* users (purple, orange and grey lines, respectively). Toxicity and Engagement were computed over 36 weeks around the ban at Week= 0, for r/The_Donald (Fig. 2a), r/GenderCritical (Fig. 2b), and r/Incels (Fig. 2c).

shown that users from banned communities may also become more toxic in other communities on the mainstream platform after the ban.

Relationship between prior and present work. We analyze how co-participation in banned antisocial communities, now hosted in less moderated spaces, i.e., “fringe” platforms, increases antisocial behavior on the mainstream platform. While previous work suggests that deplatforming may “backfire” due to creating more toxic communities on alternative platforms, we show that, additionally, antisocial behavior spills over onto mainstream platforms through co-active users.

3 Data

We use data from the three communities r/The_Donald, r/Incels, and r/GenderCritical (see Section 2 for details). 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 consisting of both submissions and comments for the users active in these communities (i) on Reddit and (ii) on the relative fringe platform.

Reddit. We collect all posts from Reddit through the Pushshift API [4]. We collect all posts made on the three focal subreddits, starting eighteen weeks before they were banned. Specifically, for r/Incels, we collected data between July 20, 2017, and November 7, 2017; for r/The_Donald, between November 11, 2019, and February 26, 2020; and for r/GenderCritical between February 14, 2020, and June 29, 2020. Overall, we collect four million posts from the three subreddits. Additionally, for each studied subreddit, we collect all contributing users’ entire

Reddit posting history. To remove users with low activity in the banned subreddit [as commonly done in social computing research, see Kumar *et al.* [23] and Samory and Mitra [41]], we consider only “focal users,” those with more than ten posts in the banned subreddit in the period prior to the banning. Finally, to filter activity on small subreddits, we remove posts made in subreddits with less than five contributions from focal users. The processed dataset contains 181,787,627 million posts made on 72,991 subreddits by 69,970 users (61,569 for r/The_Donald, 5,367 for r/GenderCritical, and 3,034 for r/Incels).

Fringe Platforms. We implement and use custom web crawlers to collect data from *thedonald.win*, *incels.co*, and *ovarit.com*, the fringe platforms where users of r/The_Donald, r/Incels, and r/GenderCritical respectively migrated following their ban. For each platform, we collect all posts made eighteen weeks before and after the ban. We collect over 2.5 million posts by 38,510 users from *thedonald.win*, 90,000 posts by 1,560 users from *ovarit.com*, and 400,000 posts by 2,270 users from *incels.co*.

Users labeling. 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 track *co-active* users across platforms by exact string-matching their usernames. Note that we assume that users with the same username across platforms correspond. A similar approach has been taken in previous work [16, 27]. Note that r/The_Donald even had a system to facilitate username continuity across platforms [10]. Finally, we filter these users, keeping only those who made at least five posts on Reddit and the fringe platform after the ban and posted on the fringe platform *only* after the ban. We obtain 1,016 Reddit users *co-active* on *thedonald.win*, 176 Reddit users *co-active* on *ovarit.com*, and 286 Reddit users *co-active* on *incels.co*.

We label all users posting on Reddit without a matching username on the fringe platform as *Reddit-only* users. We find 10,829 *Reddit-only* users that were previously members r/The_Donald, 1,228 for r/GenderCritical *Reddit-only*, and 2,753 for r/Incels.

4 Methods

To quantify the effect of co-participation on users’ behavior on Reddit, we compare *co-active* and *Reddit-only* users. We proxy antisocial behavior through users’ toxicity (as measured through Perspective API) and their activity in other extreme subreddits (controversial group engagement). To estimate the causal effects in observational data, we combine two widely used quasi-experimental causal inference methods: propensity score matching and difference-in-differences.

4.1 Propensity Score Matching

We use a one-to-one propensity score matching to match *co-active* and *Reddit-only* users that were similar in the pre-banning period. Propensity score matching (PSM) is a simple yet powerful method to account for selection bias that balances the distribution of observed covariates between groups. This method allows us to mitigate the risk that observed differences in post-banning antisocial behavior exhibited by *co-active* and *Reddit-only* users

User Characteristics	
Participation	Proportion of users' posts in the banned subreddit weighted by similarity.
Generality Score	Activity diversity [48]
First Post Time	Time of first post in the subreddit
Language Characteristics	
Toxicity	A measure for usage of toxic language
Anger and Anxiety	Frequency of anger or anxiety words
Group Characteristics	
k-core centrality	Network embedness
Eigencentality	Non-local network centrality

Table 1: Description of the covariates used in the propensity score matching to ensure that *Co-Active* and *Reddit-Only* users are comparable. See Appendix A for details

come from user characteristics, e.g., co-active users may be more toxic pre-banning and respond differently to the banning event. PSM ensures that we consider users with equal probability to become active on the fringe platform.

PSM consists of three stages: (i) propensity score modeling, (ii) propensity score matching, and (iii) estimating a treatment effect after a successful balance check. (i) We train a logistic regression classifier (LRC) to estimate the likelihood that a user will post on the fringe platform after the banning—the propensity score. In particular, we trained the LRC on a set of user features computed on the pre-banning activities described in Table 1. (ii) We match each *co-active* user to a *Reddit-only* user using the nearest neighbor algorithm. We discard matches with a similarity below 0.68. See Appendix A.1 for more details. (iii) We test the quality of the matching by measuring the standardized mean difference of each covariate used in the PSM. We obtained absolute standardized mean differences smaller than the standard 0.1 threshold [3].

4.2 Difference-in-differences

Considering the matched sample in the eighteen weeks before and after the ban date of each subreddit, we estimate the effect of co-activity in a fringe platform on users' behavior on Reddit with the following difference-in-differences (DiD) model:

$$Y_{it} = \beta_0 + \beta_1 \text{Coactive}_i + \beta_2 \text{Period}_t + \beta_3 \text{Coactive}_i \text{Period}_t + \varepsilon_{it}, \quad (1)$$

where Y_{it} is user i 's outcome (e.g., toxicity, we discuss outcomes in Section 4.3) in period t on Reddit. Coactive_i indicates if user i is *co-active* or not. Period_t indicates if the current time t is before or after the ban ($t = 0$), and ε is the error term. Under the assumption that the difference in outcomes between *co-active* and *Reddit-only* users is constant over time in the absence of co-participation on fringe platforms (the “parallel trends assumption”), the coefficient β_3 captures the causal effect of co-participation in the fringe community on the outcome variable.

4.3 Outcome Variables

Toxicity. Previous works have shown how subreddits like r/The_Donald, r/GenderCritical, and r/Incels are prone to toxic language use [16]. These subreddits are home to many antisocial behaviors, such as incivility, harassment, trolling, and cyberbullying. In this direction, the work of Grover and Mark [14] is particularly relevant, as it suggests that antisocial behaviors may be captured through automated text analysis. Therefore, we employ the Perspective API [19] to measure the toxicity level of users' posts; see Section 2 for details. To infer a user's i toxicity, we compute the median toxicity score T_{it} of all the user's posts within a given time window t . Specifically, for each user, we group their posts in weekly time windows to obtain weekly toxicity scores.

Engagement in controversial subreddits. We measure users' engagement with other controversial communities on Reddit as a second proxy for antisocial behavior. For each user i , we compute the number of posts made in subreddits hosting discussions similar to the banned subreddit during a time window t . We normalize this number by the total number of posts on the whole Reddit made by i in the same time window. We refer to the resulting measure E_{it} as engagement :

$$E_{it} = \frac{\sum_{s \in S_K} \|P_{it}^s\|}{\sum_{s \in R} \|P_{it}^s\|}, \quad (2)$$

where S_K is the set of the k -th most similar subreddits to either r/The_Donald, r/GenderCritical, and r/Incels, R is the set of all subreddits in Reddit (excluding the focal ones), and $\|P_{it}^s\|$ is the number of posts made by user i at time t in subreddit s .

To find the k -th most similar subreddits to r/The_Donald, r/GenderCritical, and r/Incels, we create a similarity scale in the interval $[-1, +1]$ where 1 represents the highest similarity to the focal subreddit (see Appendix A.1 for details). We fix $k=50$ and manually verify that the similar subreddits host indeed similar discussions to r/The_Donald, r/GenderCritical, and r/Incels.

5 Results

We combine a large-scale longitudinal and regression analyses to assess the effect of co-participation in antisocial fringe platforms on users' behavior on Reddit. We find that users co-participating on Reddit and fringe platforms exhibit increased antisocial behavior following a community ban. More importantly, our study finds that the antisocial behavior of *co-active* users *diverges over time* from that of *Reddit-only* users. We perform this analysis with two measures of antisocial behavior: (i) language toxicity and (ii) engagement with other controversial subreddits. Our results are consistent for both measures across all the studied communities.

5.1 Longitudinal Analysis

The upper row of Fig. 2 shows the toxicity of posts written on Reddit by users of r/The_Donald, r/GenderCritical, r/Incels before and after the ban. Note that here we consider the matched sample obtained after propensity score matching. Following the ban, we observe that all users increase their toxicity (Fig. 2 grey line). The post-banning average toxicity grows by up to 61% of the pre-banning toxicity (from 13 to 22). This observation confirms the finding by Trujillo and Cresci [47] of a marked increase in toxicity in the aftermath of community bans. By

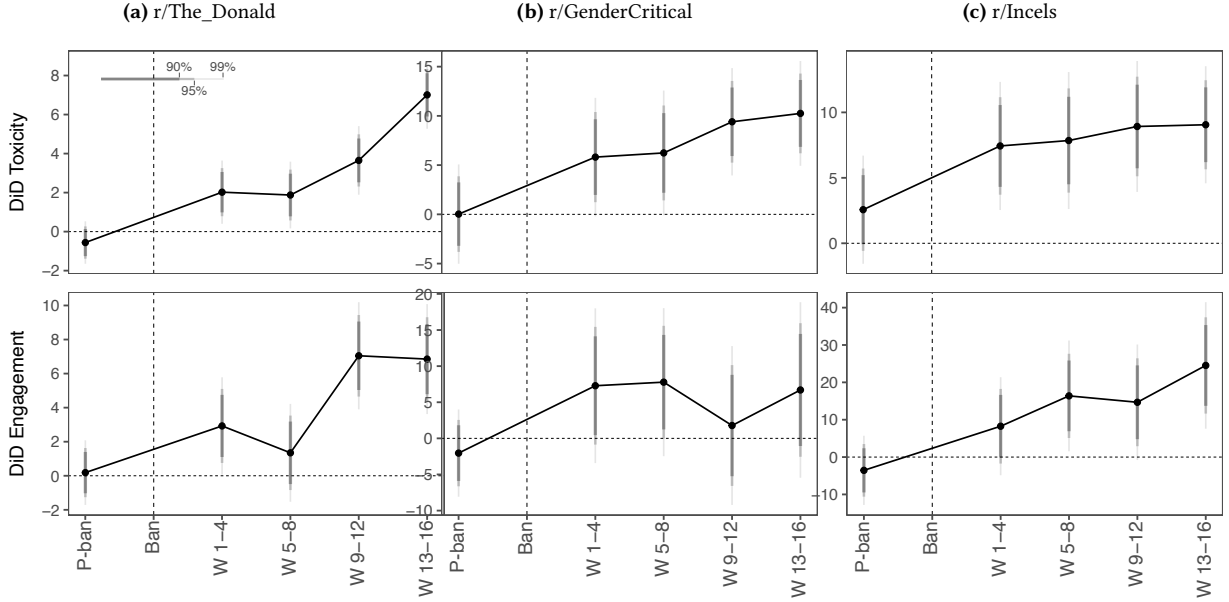


Figure 3: Estimated DiD effect of co-participation for toxicity (top row) and engagement (bottom row) shown for r/The_Donald (Fig. 3a), r/GenderCritical (Fig. 3b), and r/Incels (Fig. 3c). Effects are shown for the 5 four-weeks chunks (one for the pre-banning and four for the post-banning period). Error bars represent the 99%, 95% and 90% CIs. Errors are clustered at user level. For details see Table 2 (top)

comparing *co-active* and *Reddit-only* users separately (purple and orange lines, respectively), we see that the toxicity of *co-active* users grows faster than that of *Reddit-only*. For r/The_Donald, Fig. 2a shows a 68% average increase in *co-active* users' toxicity after the ban. This is a net increment of 23% compared to *Reddit-only* users. Similarly, we find a net increment of 41% and 20% for r/GenderCritical and r/Incels, respectively.

Qualitatively similar conclusions can be drawn when observing the engagement of *co-active* and *Reddit-only* users. For instance, in the bottom row of Fig. 2, we observe that *co-active* users of r/The_Donald and r/Incels exhibit a steady increase in engagement towards controversial communities. In particular, *co-active* users of r/The_Donald increase their engagement from 9 to 19, while in r/Incels, they go from 15 to 50. The case of r/Incels is particularly interesting, as the community remained active both on Reddit and in the subreddit r/braincels. This subreddit gave continuity to members of r/Incels as they maintained their antisocial behavior habits. However, even under these circumstances *co-active* users show more engagement towards Incels-related content than *Reddit-only* users (see Fig. 2c bottom)

5.2 Difference-in-difference Analysis

We analyze the differences highlighted above with the DiD regression introduced in Section 4. We quantify the effect of *co-participation* in fringe platforms on *Reddit* antisocial behavior. To do so, we consider the four weeks before the ban as our pre-banning reference, and we group post-banning periods into four-weeks chunks. We formalize this regression following Eq. (1). Specifically, the dependent variables Y_{it} are the toxicity (T_{it}) and the

engagement (E_{it}) of each user i at time t grouped by Period_t , i.e., four-weeks chunks. The categorical variable Period_t refers to any of the five four-weeks chunks (one pre-banning and four post-banning).

In Fig. 3, we show the DiD effect, i.e., the difference in toxicity or engagement between *co-active* and *Reddit-only* net of pre-ban differences. We observe that the pre-ban difference is zero, suggesting that the propensity score matching (see Section 4.1) adequately controls for pre-ban differences. Most importantly, from Fig. 3, we find that the DiD effects associated with each post-banning period increase over time for both toxicity and group engagement. The four DiD coefficients (reported in Table 2) increase with time, indicating that Co-Active users on Reddit become more toxic and engage more with controversial subreddits. This result provides evidence that the adoption of antisocial behaviors by *co-active* users not only increases but also *diverges* from that of *Reddit-only* users. However, we do not find evidence of such divergence in the cases of engagement for r/GenderCritical (see Fig. 3b(bottom)). We speculate this might be since r/GenderCritical was banned with other 2,000 subreddits. Such a mass ban might have caused most of the controversial communities associated with r/GenderCritical to get banned, too, thus limiting the ability of r/GenderCritical users to regroup. For instance, the two subreddits r/TrueLesbians and r/Gender_Critical, closely associated with r/GenderCritical, were jointly banned.

Interestingly, we notice that the DiD effect increases slowly for approximately eight weeks after the ban and starts to increase faster afterward. This finding is in line with the observation that users of banned subreddit may need time to become active (i.e., writing a post) on the fringe platform. Indeed, 84% of *co-active* users make their first post on the fringe platform between the eighth and twelfth week.

Additionally, we consider that co-participation may not necessarily be binary, i.e., whether a user is coactive or not. Instead, it can be interpreted as exposure to the fringe platform. We formalize this exposure as the fraction of posts made on the fringe platform over all posts made by the user across platforms, i.e., Reddit and the fringe platform. We hypothesize that increased exposure to a fringe platform increases antisocial behavior on Reddit. We then run a regression for Eq. (1) where we substitute the binary variable Coactive_i with the continuous variable indicating the user's exposure to the fringe platform. Under this setting, we run this regression for r/The_Donald to test if increased exposure leads to an increase in antisocial behavior (i.e., toxicity and engagement). We find evidence supporting this hypothesis.

In synthesis, our results provide evidence that co-participation in fringe platforms affects users' antisocial behavior on Reddit. We show that the toxic behavior of *co-active* users diverges over time from that of *Reddit-only* users. In the following section, we estimate the rate of divergence.

5.3 Divergence Analysis

We expand the regression of Eq. (1) such that it considers the following dependent variables: (i) t as a continuous variable taking values in $[-18, +18]$; (ii) Period_t a discrete variable indicating before and after ban periods; (iii) a fixed-effect u_i for each user i . We then model the dependent variable Y_{it} as the log of T_{it} and E_{it} . This transformation addresses two issues observed in the data: the skewness of the dependent variable and a non-linear increment of

antisocial behavior over time. We formalize this regression as:

$$\begin{aligned} \log(Y_{it}) = & \beta_0 + \beta_1 \text{Coactive}_i + \beta_2 \text{Period}_t + \beta_3 t + \\ & + \beta_4 \text{Coactive}_i \text{Period}_t + \beta_5 \text{Coactive}_i \text{Period}_t + \\ & + \beta_6 \text{Period}_t t + \beta_7 \text{Coactive}_i \text{Period}_t t + u_i + \varepsilon_{it}. \end{aligned} \quad (3)$$

The coefficient β_7 captures the weekly percentage increase in antisocial behaviors of *co-active* over *Reddit-only* users. Therefore, β_7 measures the divergence between the two groups. The results are reported in Table 2(bottom). In Fig. 4, we show the fitted models for the three subreddits. Figure 4a top and bottom shows the model fitted on r/The_Donald. We observe that *co-active* users diverge consistently from *Reddit-only* users in toxicity and engagement. In particular, we find that the increase in toxicity and engagement for *co-active* users exceeds that of *Reddit-only* users by 2% and 6% *per week*, respectively. In r/Incels, the results for engagement are qualitatively similar to those of r/The_Donald. In the case of r/GenderCritical, we find that the effect size on toxicity is similar to the one observed for r/The_Donald, albeit less significant. We hypothesize that the lower statistical significance results from the smaller sample size of r/GenderCritical (3, 263 samples against 50, 628). We do not find evidence of an effect of co-participation on the toxicity of r/Incels users. This last result is not surprising as users of r/Incels continued their activity on r/brainincels. r/brainincels allowed users of r/Incels to maintain their antisocial behavior, therefore mitigating the effect of the banning. Similarly, we do not find evidence of an effect of participation on the engagement of r/GenderCritical users. Again, due to the mass ban of 2020, we argue that r/GenderCritical users could not find subreddits hosting similar groups.

With this analysis, we provide 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 that of Reddit-Only users. This differential growth results in a steady divergence in antisocial behavior once co-active users start participating in the highly toxic fringe platforms.

6 Discussion

Users on fringe platforms are exposed to a more toxic environment, which may spill over onto mainstream social media. To test whether such spillover exists, we investigate if co-active users—active on both fringe platforms and mainstream social media—become more toxic after joining the fringe platform. We study controversial communities on Reddit by combining two quasi-experimental methods: propensity score matching and difference in differences.

We find that co-active users exhibit consistent and increased antisocial behavior on Reddit. This increase diverges from users of the same banned community posting only on Reddit. In particular, we find that the effect of co-participation intensifies with time and exposure to the fringe platform. The more a user is active on the fringe platform, the more they are involved in antisocial behavior on Reddit.

Our results shed light on the relations between fringe and mainstream social media. While stakeholders of mainstream social media may consider the out-migration of users exhibiting antisocial behavior to be in their best interest, assuming that their platform and the fringe platform users migrated to are independent, our study

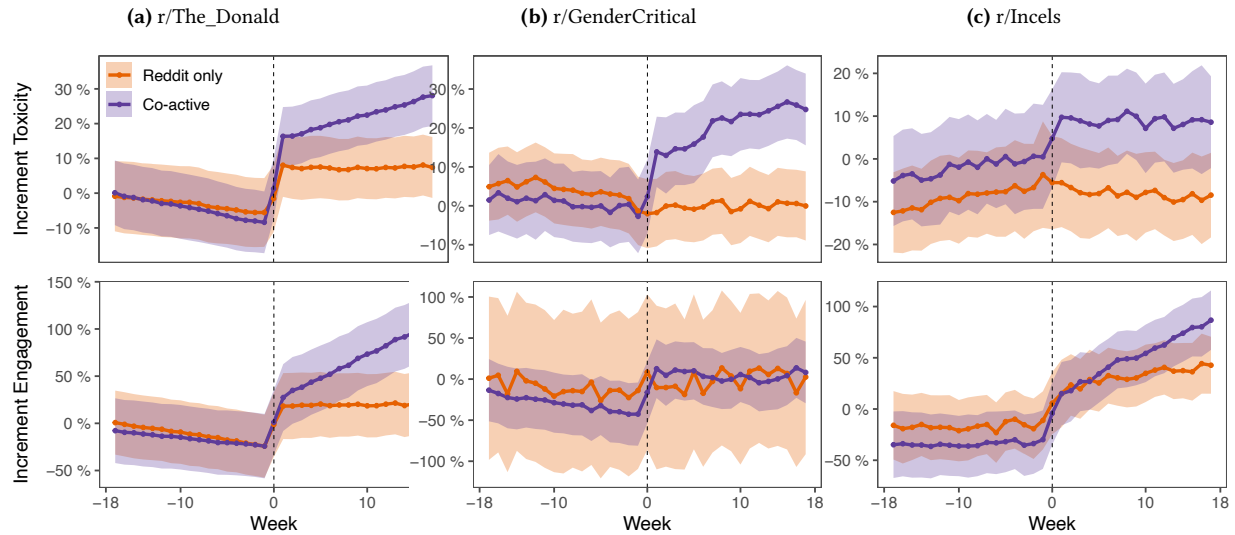


Figure 4: Divergence of toxicity (top row) and engagement (bottom row) for *co-active* (purple) and *Reddit-only* users (orange). The average predicted relative increase for toxicity and engagement are shown for r/The_Donald (Fig. 4a), r/GenderCritical (Fig. 3b), and r/Incels (Fig. 4c). The shaded areas represent the 95% CIs. For further details see Table 2 (bottom)

reveals that co-active users act as a channel through which antisocial behavior on fringe platforms spills back onto mainstream social media.

While previous work has suggested that users “adjust” to toxicity levels of existing communities on Reddit [33], our results indicate that users exposed to toxic environments on fringe platforms will act similarly on the mainstream platform. The spillover induced by co-participation should inform how administrators enact moderation policies.

Limitation and future work. Our classification of users according to their activity on fringe platforms may be inaccurate. Some users may not keep the same username or only read posts on the fringe platform. Thus, we would erroneously classify them as Reddit-Only. However, this issue does not invalidate our results. The difference-in-differences analysis we have performed yields a lower bound on the effect of co-participation on users’ antisocial behavior. This is a lower bound because perfectly labeling Reddit-only users can only increase the effect. In fact, misclassifying co-active users as Reddit-only drives up the dependent variable after the ban, thus reducing the DiD effect.

In this work, we have only broached why users become active on the fringe platform after a ban, but future work could investigate which factors play a role, e.g., push and pull factors such as the position in the social network.

7 Ethics Statement

A positive outcome of our research is that it can help mainstream platforms design policies to mitigate the spillover of antisocial behavior. For example, a platform might introduce automatic labeling of communities similar to banned ones, allowing users to make more informed decisions about their participation. However, our findings may also be used to justify turning a blind eye to problematic communities, citing spillover concerns. For example, a platform might tolerate abusive behavior in isolated communities rather than risk the spillover of that behavior to the wider platform following a ban. We primarily use publicly available data that does not require user consent. We collect data from the fringe platforms because it is an integral part of this research. We do not use any personally identifiable information (PII) from the dataset, and we do not make any inferences about individual users. Similarly, we do not name any other subreddits or users associated with the banned communities. We confirm that we have read and abide by the AAAI code of conduct.

References

- [1] Ali, S.; Saeed, M. H.; Aldreabi, E.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; Stringhini, G. (2021). Understanding the Effect of Deplatforming on Social Networks. In: *13th ACM Web Science Conference 2021*. WebSci '21, New York, NY, USA: Association for Computing Machinery, p. 187–195. ISBN 9781450383301.
- [2] Anti-Defamation League (2020). ADL Statement on Facebook's Decision to Finally Ban QAnon Content From Platform. <https://www.adl.org/news/press-releases/adl-statement-on-facebooks-decision-to-finally-ban-qanon-content-from-platform>.
- [3] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* **46**(3), 399–424.
- [4] Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; Blackburn, J. (2020). The pushshift reddit dataset. In: *Proceedings of the international AAAI conference on web and social media*. vol. 14, pp. 830–839.
- [5] Cheng, J.; Bernstein, M.; Danescu-Niculescu-Mizil, C.; Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In: *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. pp. 1217–1230.
- [6] Cheng, J.; Danescu-Niculescu-Mizil, C.; Leskovec, J. (2015). Antisocial behavior in online discussion communities. In: *Proceedings of the international aai conference on web and social media*. vol. 9, pp. 61–70.
- [7] Collins, B.; Zadrozny, B. (2020). Facebook bans QAnon across its platforms. <https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339>.
- [8] Dewey, C. (2016). Washington Post — These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them. <https://wapo.st/3A07pb1>.
- [9] Dibbell, J. (1994). A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *Ann. Surv. Am. L.* , 471.
- [10] Doggoes (2020). 'I hope if you came from T_D you reserved your reddit username even if you don't plan to useit'.
- [11] Flores-Saviaga, C. I.; Keegan, B. C.; Savage, S. (2018). Mobilizing the trump train: Understanding collective action in a political trolling community. In: *Twelfth International AAAI Conference on Web and Social Media*.
- [12] Freelon, D.; Marwick, A.; Kreiss, D. (2020). False equivalencies: Online activism from left to right. *Science* **369**(6508), 1197–1201.

- [13] Grover, A.; Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 855–864.
- [14] Grover, T.; Mark, G. (2019). Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 13, pp. 193–204.
- [15] Hoffman, B.; Ware, J.; Shapiro, E. (2020). Assessing the threat of incel violence. *Studies in Conflict & Terrorism* **43**(7), 565–587.
- [16] Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW2), 1–24.
- [17] Jaki, S.; De Smedt, T.; Gwóźdz, M.; Panchal, R.; Rossa, A.; De Pauw, G. (2019). Online hatred of women in the Incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict* **7**(2), 240–268.
- [18] Jhaver, S.; Boylston, C.; Yang, D.; Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW2), 1–30.
- [19] Jigsaw (2022). Perspective API. <https://perspectiveapi.com/>.
- [20] Juneja, P.; Rama Subramanian, D.; Mitra, T. (2020). Through the looking glass: Study of transparency in Reddit’s moderation practices. *Proceedings of the ACM on Human-Computer Interaction* **4**(GROUP), 1–35.
- [21] Kaitlyn, T. (2020). The Secret Internet of TERFs. <https://www.theatlantic.com/technology/archive/2020/12/reddit-ovarit-the-donald/617320/>.
- [22] Kim, J. W.; Guess, A.; Nyhan, B.; Reifler, J. (2021). The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication* **71**(6), 922–946.
- [23] Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* **35**(6), 1547.
- [24] Lyons, M. N. (2017). Ctrl-alt-delete: The origins and ideology of the alternative right. *Political Research Associates* **20**.
- [25] Marwick, A. E.; Caplan, R. (2018). Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies* **18**(4), 543–559.
- [26] McLroy-Young, R.; Anderson, A. (2019). From “welcome new gabbers” to the pittsburgh synagogue shooting: The evolution of gab. In: *Proceedings of the international aaai conference on web and social media*. vol. 13, pp. 651–654.
- [27] 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. In: *Tenth International AAAI Conference on Web and Social Media*.
- [28] Paudel, P.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; Stringhini, G. (2021). Soros, child sacrifices, and 5G: understanding the spread of conspiracy theories on web communities. *arXiv preprint arXiv:2111.02187*.
- [29] Pew Research (2017). The state of online harassment. <https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/>.
- [30] Phadke, S.; Samory, M.; Mitra, T. (2022). Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 16, pp. 770–781.
- [31] Porter, C. E. (2004). A typology of virtual communities: A multi-disciplinary foundation for future research. *Journal of computer-mediated communication* **10**(1), JCMC1011.
- [32] Preece, J.; Maloney-Krichmar, D.; Abras, C. (2003). History of online communities. *Encyclopedia of community* **3**(1023-1027), 86.

- [33] Rajadesingan, A.; Resnick, P.; Budak, C. (2020). Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 14, pp. 557–568.
- [34] Rajadesingan, A.; Zafarani, R.; Liu, H. (2015). Sarcasm detection on twitter: A behavioral modeling approach. In: *Proceedings of the eighth ACM international conference on web search and data mining*. pp. 97–106.
- [35] Reddit (2020). Update to our content policy. https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/.
- [36] Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; Zannettou, S. (2021). The evolution of the manosphere across the Web. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 15, pp. 196–207.
- [37] Rieger, D.; Kumpel, A. S.; Wich, M.; Kiening, T.; Groh, G. (2021). Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society* 7(4), 20563051211052906.
- [38] Romano, A. (2019). Community guidelines enforcement report. <https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report>.
- [39] Russo, G.; Gote, C.; Brandenberger, L.; Schlosser, S.; Schweitzer, F. (2022). Disentangling Active and Passive Cosponsorship in the US Congress. *arXiv preprint arXiv:2205.09674*.
- [40] Russo, G.; Hollenstein, N.; Musat, C. C.; Zhang, C. (2020). Control, Generate, Augment: A Scalable Framework for Multi-Attribute Text Generation. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 351–366.
- [41] Samory, M.; Mitra, T. (2018). Conspiracies online: User discussions in a conspiracy community following dramatic events. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 12.
- [42] Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; Smith, N. A. (2019). The risk of racial bias in hate speech detection. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. pp. 1668–1678.
- [43] Schulze, H.; Hohner, J.; Greipl, S.; Girgnhuber, M.; Desta, I.; Rieger, D. (2022). Far-right conspiracy groups on fringe platforms: a longitudinal analysis of radicalization dynamics on Telegram. *Convergence*, 13548565221104977.
- [44] Seering, J.; Wang, T.; Yoon, J.; Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms. *New Media & Society* 21(7), 1417–1443.
- [45] Sipka, A.; Hannak, A.; Urman, A. (2022). Comparing the Language of QAnon-related content on Parler, Gab, and Twitter. In: *14th ACM Web Science Conference 2022*. pp. 411–421.
- [46] Tausczik, Y. R.; Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29(1), 24–54.
- [47] Trujillo, A.; Cresci, S. (2022). Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *arXiv preprint arXiv:2201.06455*.
- [48] Waller, I.; Anderson, A. (2019). Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In: *The World Wide Web Conference*. pp. 1954–1964.
- [49] Waller, I.; Anderson, A. (2021). Quantifying social organization and political polarization in online platforms. *Nature* 600(7888), 264–268.
- [50] Williams, C. (2020). The ontological woman: A history of deauthentication, dehumanization, and violence. *The Sociological Review* 68(4), 718–734.
- [51] Wulczyn, E.; Thain, N.; Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In: *Proceedings of the 26th international*

conference on world wide web. pp. 1391–1399.

- [52] Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; Suarez-Tangil, G. (2018). On the origins of memes by means of fringe web communities. In: *Proceedings of the Internet Measurement Conference 2018*. pp. 188–202.
- [53] Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Sirivianos, M.; Stringhini, G.; Blackburn, J. (2017). The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In: *Proceedings of the 2017 internet measurement conference*. pp. 405–417.
- [54] Zannettou, S.; ElSherief, M.; Belding, E.; Nilizadeh, S.; Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. In: *12th ACM Conference on Web Science*. pp. 125–134.
- [55] Zeng, J.; Schäfer, M. S. (2021). Conceptualizing “dark platforms”. Covid-19-related conspiracy theories on 8kun and Gab. *Digital Journalism* **9**(9), 1321–1343.
- [56] Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. (2019). Graph convolutional networks: a comprehensive review. *Computational Social Networks* **6**(1), 1–23.
- [57] Zuckerman, E.; Rajendra-Nicolucci, C. (2021). Deplatforming Our Way to the Alt-Tech Ecosystem. *Knight First Amendment Institute at Columbia University, January* **11**.

A Appendix

A.1 Methodology

Subreddits similarity scale. To create a similarity scale between subreddits, we map the similarity score to $[-1, +1]$, where 1 represents the highest similarity to the considered community. To do so, we follow the method proposed by Waller and Anderson [49]. We consider our focal subreddits r/The_Donald, r/GenderCritical, and r/Incels and their polar opposites r/HillaryClinton, r/asktransgender, and r/feminists, respectively. Given a subreddit s_j , we define as relevant a subreddit $s_r^{(j)}$ where at least ten users of s_j posted at least five times. We then define a graph for a focal subreddit s_j and its polar opposite $s_r^{(j)}$ (e.g., r/Incels-r/feminists). The nodes of the graphs consist of (i) s_j and $s_r^{(j)}$, and (ii) all relevant subreddits for the two polar opposites $s_r^{(j)}$ and $s_r^{(j)}$. We draw a weighted edge between two nodes if the corresponding subreddits share at least five users. The weight corresponds to the number of users shared. Inspired by recent success of graph methods to obtain vectoral representations [39, 40, 56], we train the Node2Vec [13] algorithm on each graph to get embeddings of each subreddits of the graphs. Finally, we use the cosine similarity to obtain the similarity between our considered subreddits and those included in each graph. Using this similarity scale, we compile a list of the top *fifty* most similar subreddits to r/The_Donald, r/GenderCritical, and r/Incels.

To validate the subreddit similarity scale, we refer to the concept of convergent validity. This concept measures the correlation between our similarity scale and other measures based on the same construct. We use the only publicly available subreddit embeddings by Waller and Anderson [49] for this comparison. The embeddings from Waller and Anderson [49] are not explicitly trained towards finding similarities between specific communities. Nevertheless, they provide a general measure of subreddit similarity. We calculate Spearman’s rank-order correlation between the 1000 subreddits most similar to r/The_Donald, r/GenderCritical, and r/Incels according to our and Waller and Anderson [49] ranking. We find a significant ($p < 0.05$) moderate correlation (0.64) between the two. This result corroborates that our similarity scale successfully measures similarity to r/The_Donald, r/GenderCritical, and r/Incels. We manually analyze the top 50 subreddits on the similarity scale and confirm that they host discussions similar to r/The_Donald, r/GenderCritical, and r/Incels.

In the following, we define the covariates we use to perform the PSM to match *co-active* and *Reddit-only* users.

- **Participation:** We compute *participation* following the approach of Phadke *et al.* [30]. We define the participation of a user i at time t as $p_{it} = \frac{n_{s_j} \text{sim}(s_b, s_j)}{N_i}$. Where n_{s_j} is the number of comments made on the subreddit s_j , $\text{sim}(s_b, s_j)$ is the similarity between the embeddings of the banned subreddit s_b , (e.g., r/Incels) and s_j computed as described above. N_i is the total number of comments on Reddit of user i . p_{it} is bounded between 0 to 1, with higher scores indicating high participation in the banned subreddits discussion.
- **Generality Score:** The generality score is a measures defined by Waller and Anderson [48]. It is bounded between -1 and +1. Users with a score of +1 post in multiple and diverse subreddits. Users that have a score of -1 are instead specialists. The generality score is the average cosine similarity between the embeddings of subreddits in which a user i is active and his center of mass, weighted by the number of contributions by the community. i ’s center of mass is defined as the weighted average of the embeddings of the subreddits in which i participated.

- **First Day post:** The difference in days between the date of the first post and the banning date of the subreddit.
- **Toxicity:** We compute the weekly average toxicity of a user as described in Section 4
- **Anger and Anxiety:** A count of anger and anxiety-related words identified via LIWC [46].
- **K-Core centrality:** We build a communication network using only the banned subreddits. Nodes are users, and edges exist if a user has answered another user's post more than five times. The k-core centrality is the subgraph of nodes in the k-core but not in the (k+1)-core.
- **Eigencentality:** Using the same network we used to compute the k-core centrality, we compute the eigencentality of each node.

Outcome Variables. To compute the engagement in controversial groups for a user i , we need to individuate the top-K most similar communities to the subreddit associated with user i , i.e., `r/The_Donald`, `r/GenderCritical`, `r/Incels`. In particular, according to the similarity scale defined above, we identify the top K subreddits as those more similar to `r/The_Donald`, `r/GenderCritical`, `r/Incels`.

DiD Analysis (users-clustered std. errors)						
	r/The_Donald		r/GenderCritical		r/Incels	
	Toxicity	Engagement	Toxicity	Engagement	Toxicity	Engagement
Coactive	-0.560 (0.423)	0.185 (0.737)	0.025 (1.957)	-2.041 (2.343)	2.567 (1.604)	-3.554 (3.594)
Coactive:Period1	2.021** (0.629)	2.926** (1.110)	5.813* (2.335)	7.288* (4.156)	7.436*** (1.900)	8.254 (5.096)
Coactive:Period2	1.876** (0.665)	1.348 (1.116)	6.236* (2.458)	7.782* (3.976)	7.850*** (2.031)	16.378** (5.752)
Coactive:Period3	3.649*** (0.683)	7.046*** (1.222)	9.400*** (2.115)	1.778 (4.265)	8.923*** (1.941)	14.669* (6.001)
Coactive:Period4	7.009*** (0.674)	6.851*** (1.254)	10.250*** (2.064)	6.699 (4.714)	9.055*** (1.736)	24.527*** (6.571)
<i>Controls</i>						
(Intercept)	13.500*** (0.301)	8.854*** (0.517)	16.664*** (1.267)	9.047*** (1.156)	20.584*** (1.112)	22.250*** (2.757)
Period	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.088	0.034	0.084	0.019	0.078	0.120
Adj. R ²	0.087	0.034	0.079	0.013	0.075	0.117
Num. obs.	23158	27287	1783	1661	3129	2686
Divergence Analysis (users fixed effects)						
Coactive	-0.300 (0.193)	3.788*** (0.306)	0.095 (0.172)	-0.216 (0.623)	-0.501 (0.306)	0.646 (0.523)
Banning	0.302*** (0.018)	0.565*** (0.031)	-0.031 (0.058)	0.195 (0.129)	-0.064 (0.050)	0.494*** (0.121)
t	-0.007*** (0.001)	-0.020*** (0.002)	-0.009* (0.004)	-0.019* (0.009)	0.009* (0.004)	0.008 (0.008)
Coactive:Banning	0.328*** (0.023)	0.077* (0.043)	0.437*** (0.084)	0.302 (0.199)	0.392*** (0.072)	0.123 (0.161)
Coactive:t	-0.005*** (0.001)	0.006* (0.003)	0.006 (0.006)	0.004 (0.013)	-0.007 (0.005)	0.005 (0.011)
Banning:t	0.009*** (0.002)	0.021*** (0.003)	0.011* (0.006)	0.022* (0.013)	-0.009 (0.005)	0.023* (0.012)
Coactive:Banning:t	0.022*** (0.002)	0.060*** (0.004)	0.012 (0.009)	-0.021 (0.020)	0.004 (0.008)	0.061*** (0.015)
<i>Controls</i>						
(Intercept)	2.859*** (0.157)	-0.096 (0.215)	2.408*** (0.122)	1.123* (0.569)	2.922*** (0.239)	0.938* (0.381)
User Fixed Eff.	Yes	Yes	Yes	Yes	Yes	Yes
R ²	0.373	0.406	0.358	0.476	0.367	0.442
Adj. R ²	0.346	0.384	0.331	0.336	0.325	0.407
Num. obs.	50628	50628	3263	3263	5433	5432

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; $\cdot p < 0.1$

Table 2: Regression tables. **(top)** Coefficient estimates and clustered standard errors for the DiD analysis. **(bottom)** Coefficient estimates and standard errors for the divergence analysis. In the second regression, the response variables have been log-transformed.