Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels

MANOEL HORTA RIBEIRO, EPFL, Switzerland
SHAGUN JHAVER, Rutgers University, USA
SAVVAS ZANNETTOU, Max Planck Institute for Informatics, Germany
JEREMY BLACKBURN, Binghamton University, USA
GIANLUCA STRINGHINI, Boston University, USA
EMILIANO DE CRISTOFARO, University College London, United Kingdom
ROBERT WEST, EPFL, Switzerland

When toxic online communities on mainstream platforms face moderation measures, such as bans, they may migrate to other platforms with laxer policies or set up their own dedicated websites. Previous work suggests that *within* mainstream platforms, community-level moderation is effective in mitigating the harm caused by the moderated communities. It is, however, unclear whether these results also hold when considering the broader Web ecosystem. Do toxic communities continue to grow in terms of their user base and activity on the new platforms? Do their members become more toxic and ideologically radicalized? In this paper, we report the results of a large-scale observational study of how problematic online communities progress following community-level moderation measures. We analyze data from r/The_Donald and r/Incels, two communities that were banned from Reddit and subsequently migrated to their own standalone websites. Our results suggest that, in both cases, moderation measures significantly decreased posting activity on the new platform, reducing the number of posts, active users, and newcomers. In spite of that, users in one of the studied communities (r/The_Donald) showed increases in signals associated with toxicity and radicalization, which justifies concerns that the reduction in activity may come at the expense of a more toxic and radical community. Overall, our results paint a nuanced portrait of the consequences of community-level moderation and can inform their design and deployment.

CCS Concepts: • **Information systems → Social networking sites**; **Social networks**; • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: online communities, fringe online communities, content moderation, online radicalization, deplatforming, social networks

ACM Reference Format:
Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 316 (October 2021), 24 pages. https://doi.org/10.1145/3476057

Authors' addresses: Manoel Horta Ribeiro, EPFL, Switzerland, manoel.hortaribeiro@epfl.ch; Shagun Jhaver, Rutgers University, USA, shagun.jhaver@rutgers.edu; Savvas Zannettou, Max Planck Institute for Informatics, Germany, szannett@mpi-inf.mpg.de; Jeremy Blackburn, Binghamton University, USA, jblackbu@binghamton.edu; Gianluca Stringhini, Boston University, USA, gian@bu.edu; Emiliano De Cristofaro, University College London, United Kingdom, e.decristofaro@ucl.ac.uk; Robert West, EPFL, Switzerland, robert.west@epfl.ch.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2573-0142/2021/10-ART316 $15.00
https://doi.org/10.1145/3476057

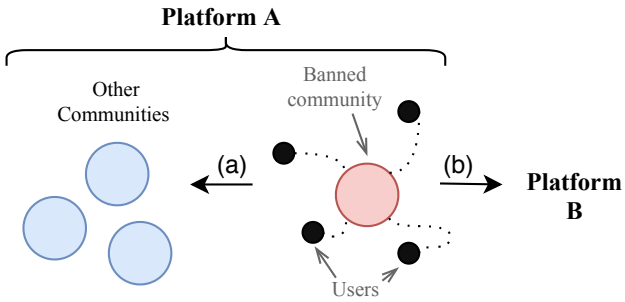

Fig. 1. **Motivation:** As a result of community-level bans, users from affected communities may choose to *(a)* participate in other communities on the same platform or *(b)* migrate to an alternative, possibly fringe platform where their behavior is considered acceptable. Scenario *a*, on which most prior work has focused, is more amenable to data-driven analysis. The present paper, on the contrary, focuses on the harder-to-analyze scenario *b*.

1 INTRODUCTION

Warning: this work quotes slur terms that some may find offensive.

The term "content moderation" is commonly associated with the process of screening the appropriateness of user-generated content, as well as imposing penalties on *users* who break the rules [44]. However, social networking platforms sometimes host entire *communities* that systematically defy regulations. There, host platforms oftentimes ban or limit the functionalities of one or several online communities. This has happened, for example, when Reddit decided that not all communities were welcome on the platform [13] and banned subreddits like `r/FatPeopleHate` and `r/transfags`. Also, more recently, following the 2021 storming of the United States Capitol, groups supporting far-right ideologies and the QAnon movement have been banned across different mainstream social media platforms [5].

The extent to which platforms should be the judges, juries, and executioners of these interventions is a topic of heated debate and has prompted experiments of governance models with societal participation [10]. There, platforms outsource some of their policy decisions—e.g., should we ban an online movement from our platform?—to a panel of experts (e.g., journalists, politicians, lawyers) representing the public interest [1].

Nonetheless, we are still left with answering a preceding question: is community-level moderation effective to begin with? We argue why this is not obvious, visually, in Fig. 1, depicting possible decisions (*a* and *b*) that users (the black dots) associated with a recently banned toxic[1] community may take. The users may *(a)* continue to be active on the same platform and participate in other groups and communities there, or *(b)* abandon the platform altogether and migrate to a different platform. In both scenarios, community-level moderation could have unintended consequences. In scenario *a*, the moderation measure could set loose an army of trolls across the platform, creating issues in other communities or *new* problematic communities [7]. In scenario *b*, the ban could unintentionally strengthen an alternative platform (e.g., 4chan or Gab) where problematic content goes largely unmoderated [31]. From the new platform, the harms inflicted by the toxic community on society could be even higher.

[1]We use 'toxic' as an umbrella term to refer to socially undesirable content: sexist, racist, homophobic, or transphobic posts, targeted harassment, and conspiracy theories that target racial or political groups.

Previous work has addressed the *"within platform"* concern. Chandrasekharan et al. [7] and Saleem and Ruths [48] studied what happened following Reddit's 2015 bans, finding that users who remained on the platform drastically decreased their usage of hate speech and that counter-actions taken by users from the banned subreddits were promptly neutralized. More broadly, Rajadesingan et al. [38] showed that, when "toxic users" migrate to healthy communities, they reduce their toxicity levels.

Nevertheless, the concern that migrations to an *alternative* platform would strengthen the toxic communities or make them more ideologically radical is still largely unexplored. Existing work suggests that, in the wake of community-level moderation, users actively seek out, and migrate to, alternative websites where their speech will not be censored [33, 50]. However, partly due to the data collection challenges posed by cross-platform studies, quantitative work on the consequences of community-level moderation *across platforms* has remained at the simulation level [24].

Present work. This paper presents an observational study of the efficacy of community-level moderation across platforms. We examine two popular communities that were originally created and grew on Reddit, r/The_Donald and r/Incels. Faced with sanctions from the platform, they created their own standalone websites—thedonald.win and incels.co—and encouraged their Reddit user base to mass migrate to the new websites. To assess whether community-level moderation measures were effective in reducing the negative impact of these communities (which we refer to as *TD* and *Incels*, respectively), we study how they progressed following their platform migrations. More specifically, we ask:

RQ1 Have the communities retained their activity levels and their capacity to attract new members following the migration to a new platform?
RQ2 Have the communities become more toxic or ideologically radical following the migration to a new platform?

Both dimensions are crucial to assess whether community-level moderation measures were truly effective. If the communities simply "changed addresses" and grew larger and more toxic on the new platforms, the moderation measures may have actually increased their capacity to harm society as well as their own members; e.g., outside of Reddit, these communities might orchestrate online harassment campaigns more effectively or disseminate more hate speech.

Materials and methods. To study how migrations affect communities, we leverage over 6 million posts made by more than 138 thousand users pooled across the platforms before (Reddit) and after (standalone websites) the migration event. We extract activity-related signals, such as the number of posts, active users, and newcomers, as well as content-related signals, such as algorithmically derived "toxicity scores," that aim to identify behaviors indicative of user radicalization, such as fixation and group identification [11]. Employing quasi-experimental setups, including matching and regression discontinuity analysis, we study these signals from a *community-level perspective*, analyzing how daily activity and overall content changed, and from a *user-level perspective*, examining how the behavior of individual users changed following platform migrations.

Summary of findings. Analyzing activity levels and the inflow of newcomers to the communities (**RQ1**), we find that the moderation measures significantly reduced the overall number of active users, newcomers, and posts in the new communities compared to the original ones. However, individually, users posted more often on the alternative platforms. A closer look at the users whom we managed to match before vs. after the migration suggests that this increase in *relative activity* is due more to self-selection rather than behavior change. Users who migrated were more active in the original platform, and their activity dropped on a user level.

Analyzing changes in the content being posted in the communities following the migration (**RQ2**), we find evidence that users in the TD community became more toxic, negative, and hostile when talking about "objects of fixation" (e.g., democrats, leftists). Changes in the usage of third-person plural (e.g., "they") and first-person plural (e.g., "we") pronouns also indicate an increase in ingroup identification and in othering language. For the Incel community, we find that changes tend to be statistically non-significant.

Implications. Our analysis suggests that community-level moderation measures decrease the capacity of toxic communities to retain their activity levels and attract new members, but that this may come at the expense of making these communities more toxic and ideologically radical. Therefore, as platforms moderate, they should consider their impact not only on their own websites and services, but in the context of the Web as a whole. Toxic communities respect no platform boundary, and thus, platforms should consider being more proactive in identifying and sanctioning toxic communities before they have the critical mass to migrate to a standalone website. Overall, we expect that our nuanced analysis will aid stakeholders to take moderation decisions and make moderation policies in an evidence-based fashion.

2 BACKGROUND AND RELATED WORK

2.1 Community-level moderation on Reddit

Reddit employs two community-wide moderation measures: *quarantining* and *banning*. When a community is quarantined, it stops appearing in Reddit's search results and front page. Moreover, users who attempt to access quarantined subreddits (directly through their URLs) are met with a splash page warning them of the shocking or offensive content contained inside. In contrast, banning a community makes it inaccessible and removes all its prior posts. Quarantining frequently precedes banning, so in practice, it serves as a warning to the subreddit to reform itself.

The history of community-level moderation in Reddit dates back to 2015 when Reddit banned five subreddits for infringing their anti-harassment policy [13]. Newell et al. [33] studied how these bans led users to migrate towards alternative platforms (e.g., Voat). Using a mix of self-reported statements and large-scale data analysis, they identified reasons why users left Reddit and found that alternative platforms struggled to attain the same diversity of communities as Reddit. The effects of these bans *within* Reddit were also extensively studied [7, 48], as previously discussed. Overall, findings from these studies suggest that the bans worked for Reddit: they led to sustained reduced interaction of users with the Reddit platform; users who stayed became less toxic after they migrated to other communities within Reddit; and counter-actions taken by users (e.g., creating alternative subreddits) were not effective.

2.2 Communities of interest

TD. The r/The_Donald subreddit (TD) was created on 27 June 2015 to support the then presidential candidate Donald Trump in his bid for the 2016 U.S. Presidential election. The discussion board, linked with the rise of the Alt-right movement at large, has been denounced as racist, sexist, and islamophobic [29]. Its members often engaged in "political trolling," harassing Trump's opponents, promoting satirical hashtags, and creating memes with pro-Trump and anti-Clinton propaganda [14]. TD is also known for spreading unsubstantiated conspiracy theories like Pizzagate [34] and the Seth Rich murder conspiracy [16].

We depict important events in TD's history in Fig. 2. The subreddit was quarantined in mid-June 2019 for violent comments, and on 26 February 2020, Reddit administrators removed a number of TD's moderators, and the community was placed under a "restricted mode," which removed the

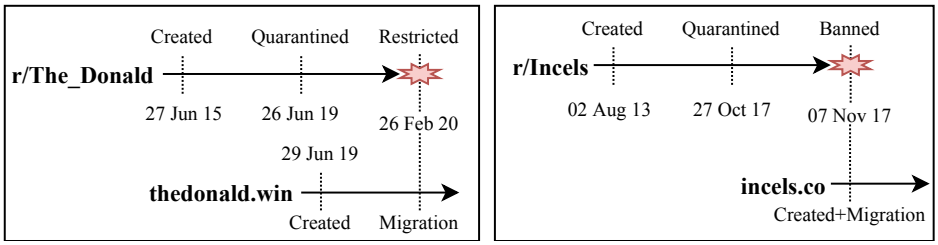

Fig. 2. **Timelines**: We depict the dates of creation, quarantining, and banning for the two communities studied here.

ability of most of its users to post. Months after the subreddit became inactive, it was banned in late June 2020. While these moderation measures were taking place, TD users were actively organizing a "plan B." In 2017, its members were already considering migrating to alternative platforms [45], and in 2019, after getting quarantined, moderators created a backup site, thedonald.win, that was promoted in the subreddit using stickied posts [47] (i.e., always shown among the first in the feed for the community). TD users continued using the subreddit until the community became "restricted." Then, they largely flocked to the alternative website [55]. Note that, although TD was eventually banned, we focus here on its "restriction," since it was this measure that halted user participation and ignited the community migration.

Incels. The r/Incels subreddit was created in August 2013. Short for involuntary celibates, it was a community built around "The Black Pill," the idea that looks play a disproportionate role in finding a relationship and that men who do not conform to beauty standards are doomed to rejection and loneliness [17, 28]. Incels rose to the mainstream due to their association with mass murderers [21] and their obsession with plastic surgery [20]. The community has been linked to a broader set of movements [28, 40] referred to as the "Manosphere," which espouses anti-feminist ideals and sees a "crisis in masculinity." In this world view, men and not women are systematically oppressed by modern society. Lately, specialists have also suggested that these communities may play an important role in radicalizing disenfranchised men and producing ideological echo chambers that promote violent rhetoric [21].

The r/Incels subreddit grew swiftly in early 2017, reaching over 3,000 daily posts [40]. Shortly after, in late October 2017, it was quarantined and then banned two weeks later [51]. In an interview for a podcast [25], one of the subreddits' former core members, *seargentincel*, mentions that he had already discussed moving the community outside of Reddit with moderators. According to him, when the subreddit was banned, he created the standalone website incels.co, and former r/Incels members quickly organized the migration in Discord channels. Again, we provide exact dates for relevant events in Fig. 2.

Choice of communities. We study these two communities for two main reasons. First, due to their importance: they have a large number of members and have impacted society at large, e.g., w.r.t. conspiracy theories [16] and real-world violence [21]. Second, these are communities whose migrations were backed by community leaders, and that migrated to other public websites. Had the members of these communities spread to a loosely connected network of private channels (e.g., on Telegram), there would be several additional technical and ethical research challenges.

2.3 Toxicity and radicalization online

Internet platforms experience a myriad of toxic behaviors such as incivility [4], harassment [3, 23], trolling [9] and cyberbullying [26]. In recent years, researchers have explored the dynamics of such behaviors online aided by automatic methods [31, 40]. Broadly, the methods employed fall under one of two categories. They either *(a)* count hate-related or toxicity-related words (e.g., using HateBase [19]); or *(b)* deploy machine-learning based methods to classify comments as toxic or as hateful (e.g., Google's Perspective API [36]). Methods differ in what they intend to measure: some aim to measure "hate speech," while others "toxicity." While these concepts differ tremendously, research has suggested that measuring hate speech through text is difficult due to its contextual nature, and that machine learning classifiers struggle to distinguish between offensive and hateful speech [12, 41].

Intertwined with online toxicity are movements and ideologies that engage in harassment campaigns and real-world violence, as well as espouse hateful views towards minorities [27, 30]. Social networks have been identified as places where individuals are exposed and eventually adhere to such fringe movements [42]. In this direction, the work of Grover and Mark [18], also on Reddit, is particularly relevant, as their work suggests that behaviors indicative of radicalization such as fixation and group identification may be captured through automated text analysis. We extend their methodology to assess the changes in user-generated *content* following the migrations, using the same word categories (derived from Linguistic Inquiry and Word Count, or LIWC [35]) and developing, for the Incel and TD communities, custom-built "fixation dictionaries" that contain terms serving as objects of fixation in the communities (e.g. *leftist* for TD, *feminism* for Incels). Additionally, we use the Perspective API to measure how toxicity in these communities changed post-migration.

The suitability of using models from the Perspective API as toxicity sensors has been explored in previous work. Rajadesingan et al. [38] found that, for Reddit political communities, the performance of the classifier is similar to that of a human annotator, while Zannettou et al. [56] found that Perspective's "Severe Toxicity" model outperforms alternatives like HateSonar [12]. Perspective has been shown to be biased against comments mentioning marginalized subgroups and for comments posted in African American English [49]. We find no compelling reason to believe that these biases may impact the post-migration changes in the toxicity of the communities studied.

Lastly, it is worth stressing that the utility of understanding toxicity in online communities goes beyond the study of fringe or troublesome communities. In the context of peer production communities, Carillo and Marsany [6] have discussed toxicity drawing notions from ecology and toxicology: exposure to "toxic" content without the appropriate "defense mechanisms" would harm the productivity of online communities. In this paradigm, efforts to understand, pro-actively detect, and quickly act against antisocial or toxic behavior would be key in maintaining healthy online communities, directions that have been empirically explored by previous work [8, 39, 57].

2.4 Relation with prior work

Overall, previous research discussed above has examined the efficacy of community-level moderation *within* Reddit [7, 48] and analyzed cross-platform migrations that ensued [33]. Our work takes a significant step further, by assessing the efficacy of these interventions in a new direction. Given that communities *do* migrate following moderation measures, we study if these measures are effective when considering the development of the communities outside of their original platform. To do so, we draw from a rich literature of existing work on online toxicity and its relationship with behaviors indicative of radicalization [18], as well as on previous studies analyzing the communities at hand [14, 40].

Table 1. Overview of our datasets.

Platform	Community	Submissions	Comments	Users
Reddit	/r/Incels	17,403	340,650	18,088
	/r/The_Donald	251,090	2,703,615	80,002
Websites	Incels.co	25,138	385,765	2,270
	thedonald.win	280,156	2,390,641	38,510

3 MATERIALS AND METHODS

3.1 Data collection

We collect data from both Reddit (for the period *before* migrations) and standalone websites (for the period *after*).

Reddit. To collect Reddit data, we use Pushshift [2], a service that performs large-scale Reddit crawls. We collect all submissions and comments made on r/The_Donald and r/Incels, starting from 120 days before the moderation measure, and until its date. Specifically, for r/Incels, we collect data between 10 July 2017 and 7 November 2017; for r/The_Donald, between 29 October 2019 and 26 February 2020. Overall, we collect around 3 million comments in 260K submissions (or "threads") from both subreddits (see Table 1).

Standalone websites. We additionally implement and use custom Web crawlers to collect data from the standalone websites (incels.co and thedonald.win). For each, we collect all submissions and comments posted for a period of 120 days after the community-level moderation measure. Specifically, for incels.co, we collect data between 7 November 2017 and 6 March 2018; for thedonald.win, between 26 February 2020 and 24 June 2020. Overall, we collect over 2.5 million comments and submissions from thedonald.win and over 400K comments and submissions from incels.co. In the rest of the paper, to ease presentation, we refer to both submissions and comments as "posts."

3.2 User analysis

We briefly describe our methods for matching users across platforms and for analyzing newcomers.

Matched Users. To better understand changes at the user-level, we also carry out analyses with matched users, finding pairs of users with the exact same username on both Reddit and the standalone websites. We consider that these users are the same individuals in the two platforms, an assumption backed by anecdotal evidence from within the communities (thedonald.win even had a feature to reserve your Reddit username [54]) and by previous research [33]. Allowing for upper/lower-case differences, using this method, we were able to match 8,651 users between r/The_Donald and thedonald.win (around 20% of the user base of the latter) and 286 users between r/Incels and incels.co (around 13%).

Newcomers. We estimate the inflow of newcomers in each community considering both the pre- and post-migration period by counting the daily number of posts with usernames never before observed. For instance, if a user X posted in thedonald.win on the 1st of March of 2020, and no user with such username posted before in either thedonald.win or r/The_Donald, we would consider him or her a newcomer. Note that here, if we used only the data from 120 days before and after the migration, we would observe a spike in newcomers at the beginning of the study period. To prevent that, for each community, we additionally download all history available in Pushshift to act as a buffer.

Table 2. Fixation dictionaries.

Incels	female(s) normie(s) chad(s) virgin whore(s) girl(s) rope gf girlfriend women beta cunt suicide pussy woman bitch(es) cuck(s) feminism
TD	trans commie dem(s) democrat(s) deep communist diversity leftist communism antifa socialist left socialism libs gender

3.3 Content analysis

To understand the impact of platform migration on the content being produced by the communities, we use text-based signals associated with toxicity and user radicalization [18, 31].

Fixation dictionary. We generate a *fixation dictionary* for each of the communities, selecting terms related to their "objects of fixation." More specifically, we: (1) select terms that are more likely to occur in the communities of interest as compared to Reddit in general, and (2) manually curate these terms, selecting those that are related to these communities' objects of fixation (e.g., *women* and *feminism* for Incels). To obtain the list of terms, we extract words from the communities of interest and from a 1% random sample of Reddit for a period of one month (immediately prior to the study period previously described). We exclude bot-related messages (e.g., auto-moderation), stop-words, and words that occurred fewer than 50 times, and calculate the log-ratio between the frequency of a keyword in the communities being studied and on Reddit in general. From this, we obtain, for each community, the 250 terms that have the highest relative occurrence. Then, to build the fixation dictionary, three authors of this paper (all familiar with the communities at hand) discussed each term and came to an agreement on whether or not that term was an object of fixation. Table 2 reports the terms in our fixation dictionary for each community; be advised that the terminology in this table is offensive.

Toxicity score. To analyze content toxicity, we use Google's Perspective API [36], an API consisting of machine learning models trained on manually annotated corpora of text. More specifically, we employ the "Severe Toxicity" model which allows us to assess how likely (on a scale between 0 and 1) a post is to be "rude, disrespectful, or unreasonable and is likely to make you leave a discussion." This model is also trained specifically to not classify benign usage of foul language as toxic.

LIWC. We measure changes in word choice using the Linguistic Inquiry and Word Count (LIWC) tool [52]. LIWC consists of various dictionaries (in total 4.5K words) to classify words into over 70 categories, including general characteristics of posts (e.g., word count), linguistic components (e.g., adverbs), psychological processes (e.g., cognitive processes), and non-psychological processes (e.g., pronouns). In this work, we study changes for the following (aggregated) LIWC 2015 categories: (1) Negative Emotions: sum of the *Anger*, *Anxiety*, and *Sadness* LIWC categories; (2) Hostility: sum of *Anger*, *Swear*, and *Sexual* LIWC categories. (3) Pronouns: we focus on the usage of third-person plural (e.g., "they"), and first person plural (e.g., "we") pronouns.

Mapping signals to warning behaviors. These different signals, as well as their combinations, have been described as warning behaviors of ideological radicalization. We focus on two warning behaviors described by Cohen et al. [11]: (1) *Fixation:* a pathological preoccupation with a person or cause that is increasingly expressed with negative and angry undertones; and (2) *Group Identification:* strong identification and moral commitment to the ingroup and distancing from the outgroup. To study changes in *fixation*, we analyze our fixation dictionaries along with *Toxicity scores* and the word categories *Negative Emotions* and *Hostility*. To study changes in *group identification*, we study changes in the usage of pronouns as measured by LIWC. These choices were motivated, as discussed earlier, by previous work by Grover and Mark [18].

3.4 Ethics and reproducibility

In this work, we only used data publicly posted on the Web and did not (1) interact with online users in any way, nor (2) simulate any logged-in activity on Reddit or the other platforms. When we matched users on Reddit and the fringe platforms, we did not attempt to gain any information about users' personal identities. Anonymized reproducibility data and code are available at doi.org/10.5281/zenodo.5171068 We stress that the data is provided without the usernames or the actual text posted (i.e., only the signals extracted). We believe that this makes de-anonymization harder than crawling the standalone websites and downloading existing Reddit dumps. These steps follow previous work studying toxic communities on Reddit [38], and we believe they minimize the potential harms associated while ensuring the study is reproducible. Additionally, we note that we only do *exact matching* on publicly available data, while not singling out any individual user, and thus we believe we are not infringing on reasonable privacy expectations.

4 CHANGES IN ACTIVITY LEVELS

In this section, we measure how the community-level moderation measures changed posting activity levels and the capacity of the two communities to attract newcomers (**RQ1**). We do so from two different perspectives. First, we aggregate our data on a daily basis, inspecting *community-level* changes in the number of posts, active users and newcomers. Next, we zoom in to the *user-level* and examine how individual users' behavior changed post-migration.

4.1 Community-level trends

Fig. 3 shows the daily number of newcomers, posts, and active users in each community before and after the migrations for both the TD and the Incel community. Note that we consider data from both the subreddit and the fringe platform users migrated towards.

To gain a better understanding of the overall trends, we perform a regression discontinuity analysis for each statistic in each community. We employ a linear model:

$$y_t = \alpha_0 + \beta_0 t + \alpha i_t + \beta i_t t, \tag{1}$$

where t is the date, which takes values between -120 and $+120$ and equals 0 in the day of the moderation measure; y_t is statistic we are modeling; and i_t is an indicator variable equal to 1 for days following the moderation measure (i.e., $t > 0$), and 0 otherwise. Our model assumes that daily activity levels (for the different metrics) can be approximated by a line (defined by coefficients α_0 and β_0), which, post-migration, can change both its intercept (α) and its slope (β). We analyze these changes to understand the impact of platform migrations on the communities at hand.

We exclude data from a "grace period" of 15 days before and after the moderation measure.[2] This accounts for the bursty behavior happening to user activity metrics in the days around the migration. For example, for newcomers, many of the users who migrated to the new website (thedonald.win or incels.co) choose new usernames, which creates a spike in the metric. However, this initial spike is not interesting to capture the overall trend of newcomers in the website, and the grace period addresses that. Additionally, there were a few days on which the Pushshift ingest had problems or where there was a large volume of spam-like content. The values for the statistics on these dates are depicted as gray crosses in Fig. 3 and were not considered to fit the models. Both coefficients and 95% CI for each parameter in the regressions are shown in Table 3.

Newcomers. The first row of Fig. 3 shows the number of daily newcomers in each community (as described in Sec. 3.2). We find that, for both communities, there was a significant *decrease* in the

[2]We stress that our results are robust to changes in this parameter: we have experimented with different window sizes (e.g. 7 and 21 days), obtaining largely the same results

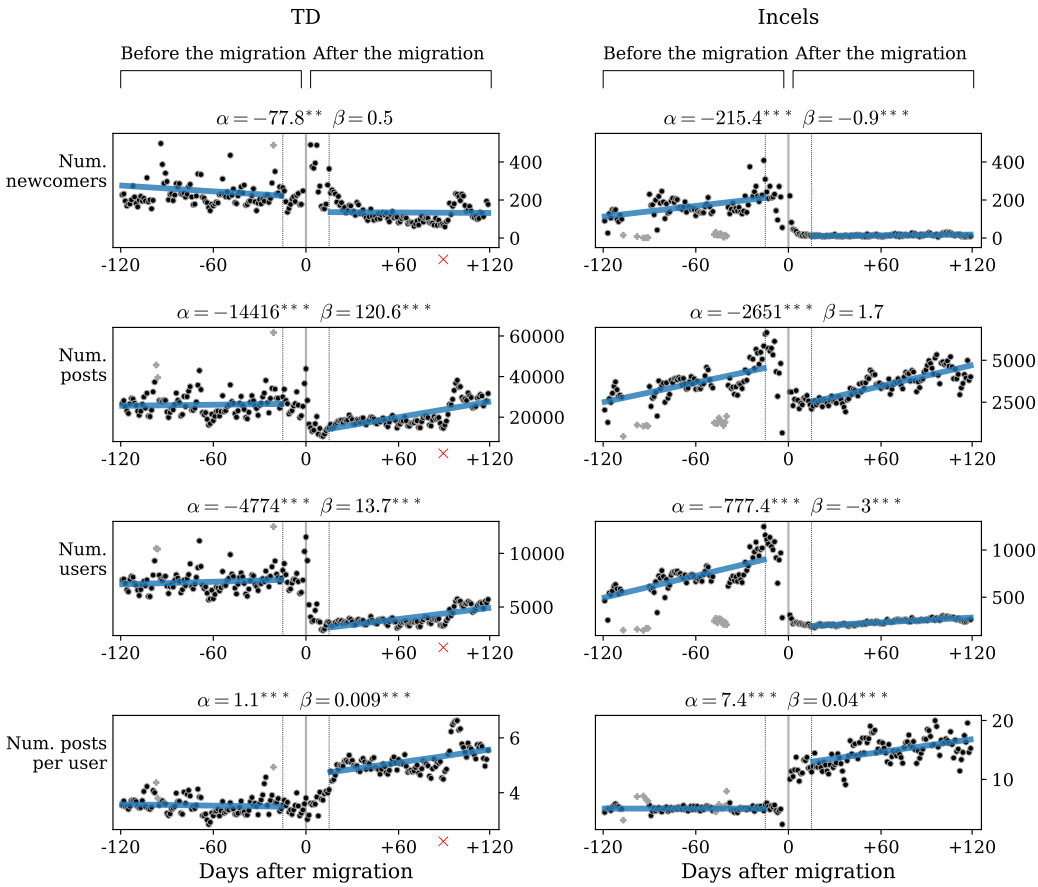

Fig. 3. **Activity levels:** Daily activity statistics for the TD community (left) and the Incel community (right) 120 days before and after migrations. Dots represent the daily average for each statistic, and the blue lines depict the model fitted in the regression discontinuity analysis. The migration date and a grace period around it (used in the model) are depicted as solid and dashed gray lines, respectively. Dots shown in gray represent days where Pushshift ingest had issues, or where there was a large volume of spam-like content. On top of each subplot, we report the coefficients associated with the moderation measure in the model (α and β). Coefficients for which $p < 0.001$, 0.01, and 0.05 are marked with ***, **, and *, respectively. For the TD community, we mark the killing of George Floyd (on 25 May 2020), with a red cross (×) close to the x-axis.

influx of newcomers following the migration. The TD community saw a significant decrease in the offset of around 78 daily newcomers ($\alpha = -77.8$). This represents a percent change of around −30% of the *M*ean *V*alue *B*efore the community-level *I*ntervention (referred to as *MVBI* henceforth), i.e., the drop represents roughly 30% of the average daily value in the pre-migration period. The decrease was even more substantial for the Incel community, which experienced around 215 fewer newcomers a day ($\alpha = -215.4$), roughly −150% of the *MVBI* (note that the drop was, therefore, *bigger* than the pre-migration average). Furthermore, the Incel community had a significant increasing trend before the migration ($\beta_0 = 0.9$), which was weakened in the post-migration period ($\beta = -0.9$).

Posts and users. The second and third rows of Fig. 3 show that both the total number of daily posts and daily posting users dropped significantly post-migration. TD experienced a decrease of around 14.4k daily posts ($\alpha = -14416$, -55% of the *MVBI*) and of around 4.7k daily active users ($\alpha = -4774$, -65% of the *MVBI*). In both cases, the slope became steeper after the migration, with a significant increase of around 121 new posts a day ($\beta = 120.6$), and around 14 additional active users a day ($\beta = 13.7$). A possible explanation for this increase is that the killing of George Floyd (25 May 2020) and the demonstrations that ensued may have boosted participation on the platform, since the date coincides with a sharp rise in both statistics. Repeating the regression analysis excluding the period after 24 May 2020, we find non-significant *decreases* in the slope (β) for both statistics, which strengthens this hypothesis. We further discuss this confounder in Sec. 6.

For the Incel community, there were significant decreases of around 2.6k posts a day ($\alpha = -2651$, -73% of the *MVBI*), and of around 777 daily active users ($\alpha = -777.4$, -116% of the *MVBI*). Looking at the trends for the number of active users, we find a significant positive trend across the whole period ($\beta_0 = 3.9$, see Table 3) but the slope decreases significantly after the migration ($\beta = -3$).

Posts per user. The fourth row of Fig. 3 shows the daily average of the posts per user ratio. Here, we find that the moderation measure *significantly increased* relative activity. The TD community showed an increase in the number of daily posts per user of around 1 extra posts per user ($\alpha = 1.1$, 31% of the *MVBI*); for Incels, the increase was of around 7 extra posts per user ($\alpha = 7.4$, 123% of the *MVBI*). In both cases there is also a significant increase in the trend ($\beta = 0.009$ for TD, and $\beta = 0.04$ for Incel). This adds nuance to the overall scenario: although the activity in the communities is reduced, *relative* to the number of users, it increases.

4.2 User-level trends

The analyses done so far paint a comprehensive picture of the changes in activity due to migration at the community-level. Yet, they do not disentangle the effects happening at the user-level. We found that *relative* activity increased (i.e., fewer users posted more often), but the underlying mechanism for this change is still unclear. Users' activity may have indeed increased *after* the migration, i.e., individually, each user who migrated might post more often on the fringe website, but the increase could also be due to self-selection: users who migrated following the moderation measure might have been more active to begin with.

Understanding the reason behind this (relative) activity increase is important in order to evaluate the efficacy of the moderation measure. If the increase occurred because users became more active, the subset of users "ignited" by the moderation measure could cause even greater harm in the new platform. However, if the increase was only due to self-selection, we might consider the measure successful in decreasing the activity and reach of the communities. To better understand the mechanism behind this activity increase, we perform an additional set of analyses inspecting what changed at the *user level* post-migration. To do so, we analyze the set of users before and after the migration, and additionally, the set of *matched users* described in Sec. 3.2.

Comparing posts-per-user distributions. We begin by comparing the distribution of posts from matched users and the general population of users both before and after the migration. Fig. 4 depicts the complementary cumulative distribution function (CCDF) of the number of posts for all users (solid line) and for matched users (dashed line) both in the fringe communities (in red) and on Reddit (in blue).

Considering all users (solid lines), the CCDFs confirm our previous analysis, showing that users are more active in the fringe websites, since the red solid line is consistently above the blue solid line. This is also captured by the mean of user activity in fringe communities, which is of 73 posts

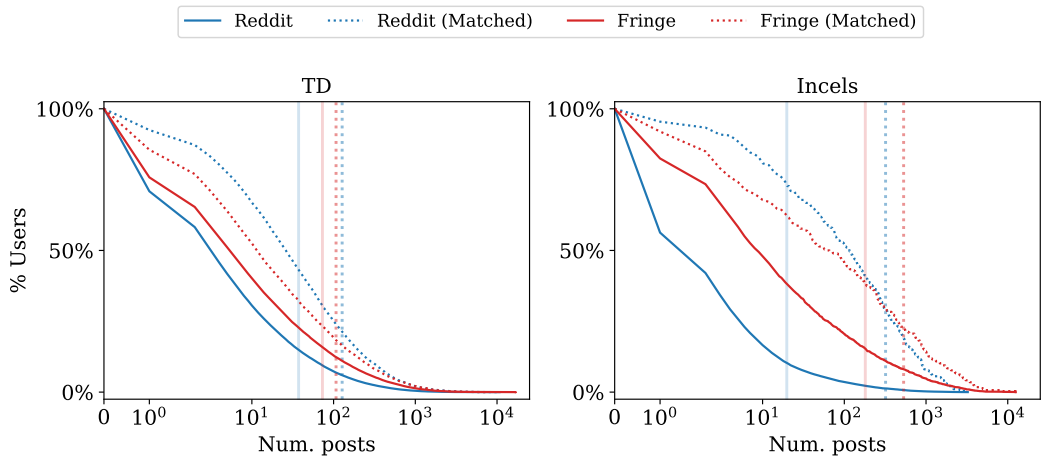

Fig. 4. **CCDFs of posts per user:** For each community, we depict the complementary cumulative distribution function (CCDF) of the number of posts per user for: (1) all users who posted in the 120 days *before* (solid blue) and *after* the moderation measure (solid red), and (2) users we managed to match based on username while they were still on Reddit (dashed blue) and on the fringe platform (dashed red). The plot also depicts the mean value for each one of these populations as vertical lines in the same color/style scheme. Recall that the CCDF maps every value in the x-axis to the percentage of values in a sample that are bigger than x (in the y-axis).

per user (95% CI: [70, 76.3][3]) for the TD community, and of 180.6 (95% CI: [155.1, 209.3]) for Incels. These values are significantly higher than in Reddit, where there are, on average, 37.3 posts per user (95% CI: [36.2, 38.5]) in the TD community, and 19.8 (95% CI: [18.2, 21.5]) in the Incel community.

Comparing the number of posts per user on Reddit in general (blue line) with users we managed to match (dashed blue line), we find that matched users are more active than users in general. In Reddit, matched users had an average of 127.3 posts (95% CI: [120.2, 135]) in the TD community, and 319.9 posts (95% CI: [264.7, 381.8]) in the Incel community, significantly higher than the average user in Reddit in each community (reported above).

Matched comparisons. The above analysis suggests that users who migrated were more active than average on Reddit, which could lead to an increase in *relative* activity due to self-selection. To further investigate this, we compare, for each matched user, the change in number of posts before and after the migration. More specifically, we analyze the log-ratio of posts before vs. after the migration for each matched user, defined as $\log_2 \frac{\text{\# posts after}}{\text{\# posts before}}$. Note that this metric provides an intuitive interpretation of the change in activity for a user: if the numbers of posts before and after the migration are the same, the log-ratio will be 0; if the user posted twice as much, it will be 1; and if the user posted half as much, -1.

In Fig. 5, we depict the mean value of the log-ratios for all users in the first column and, in the next four columns, for users stratified by their activity in the pre-migration period. We divide users in quartiles according to how much they posted in the pre-migration period[4] and then report the mean for each quartile.

[3]Confidence intervals were calculated through bootstrapping.

[4]For the TD community, the quartile ranges for the number of posts before the migration were $Q1 = [1, 7)$, $Q2 = [7, 27)$, $Q3 = [27, 101)$, $Q4 = [101, \infty)$; for Incels, $Q1 = [1, 18)$, $Q2 = [18, 116)$, $Q3 = [116, 389)$, $Q4 = [389, \infty)$.

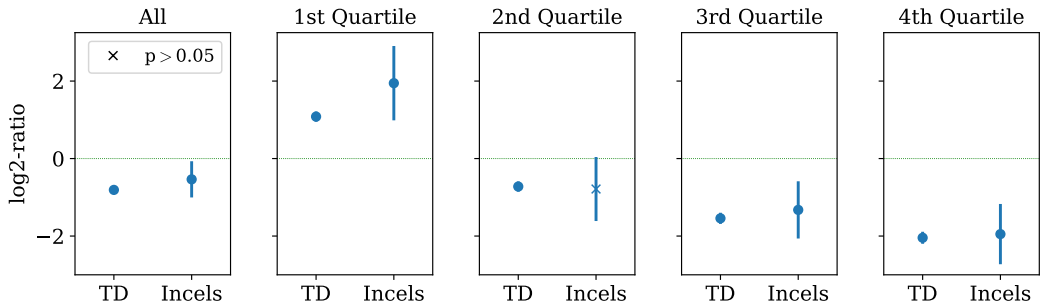

Fig. 5. **User-level change in number of posts:** Mean log-ratios between the number of posts before and after the migration for each user. In the first column, the mean is calculated for all users, while for the last four, we stratify users according to their level of pre-migration activity. The horizontal line depicts the scenario where the number of posts remained the same (log-ratio = 0). Error bars represent 95% CIs.

Considering the complete set of matched users (first column of Fig. 5), we find that the mean activity log-ratios are significantly smaller than zero for both communities: -0.81 (95% CI: [-0.86, -0.75]) for the TD community and -0.53 (95% CI: [-0.96, -0.10]) for the Incel community. This result provides further evidence for the self-selection hypothesis: not only did we find the group of matched users to be more active, but, within this group, activity has *decreased*.

Analyzing the users stratified by their activity (in the last four columns of Fig. 5), we find that this decrease in activity is stronger for users who were the most active in the pre-migration period. The mean log-ratios for each quartile in TD are, respectively, $\mu_{Q1} = 1.1$, $\mu_{Q2} = -0.7$, $\mu_{Q3} = -1.5$, and $\mu_{Q4} = -2.0$. This shows that users in the least active quartile (Q1) became around twice ($2^{1.1}$) as active, while those in the most active quartile (Q4) decreased their activity to around one-quarter ($2^{-2.0}$). For the Incel community we observe a similar pattern, with mean log-ratios of $\mu_{Q1} = 1.9$, $\mu_{Q2} = -0.8$, $\mu_{Q3} = -1.6$, and $\mu_{Q4} = -2.0$. Overall, these findings mitigate the concern that a core group of extremely dedicated users was "ignited" by the migration.

4.3 Take-aways

Our analysis suggests that community-level moderation measures significantly hamper activity and growth in the communities we study. For both communities, there was a substantial decrease in the number of newcomers, active users, and posts after the moderation measure. Yet, this tells only part of the story: we also find an increase in the *relative* activity for both communities: per user, substantially more daily posts occurred on the fringe websites.

A closer look into *user-level* indicates that this relative increase in activity is due to self-selection, rather than an increase in user activity post-migration. Not only do we find that users we managed to match were more active on Reddit before the migration, but they even *reduced* their overall activity after they went to the new platform.

5 CHANGES IN CONTENT

In this section, we use the signals described in Sec. 3.3 to analyze whether the communities and their users became more toxic and ideologically radical following the migrations. We again analyze community- and user-level trends separately.

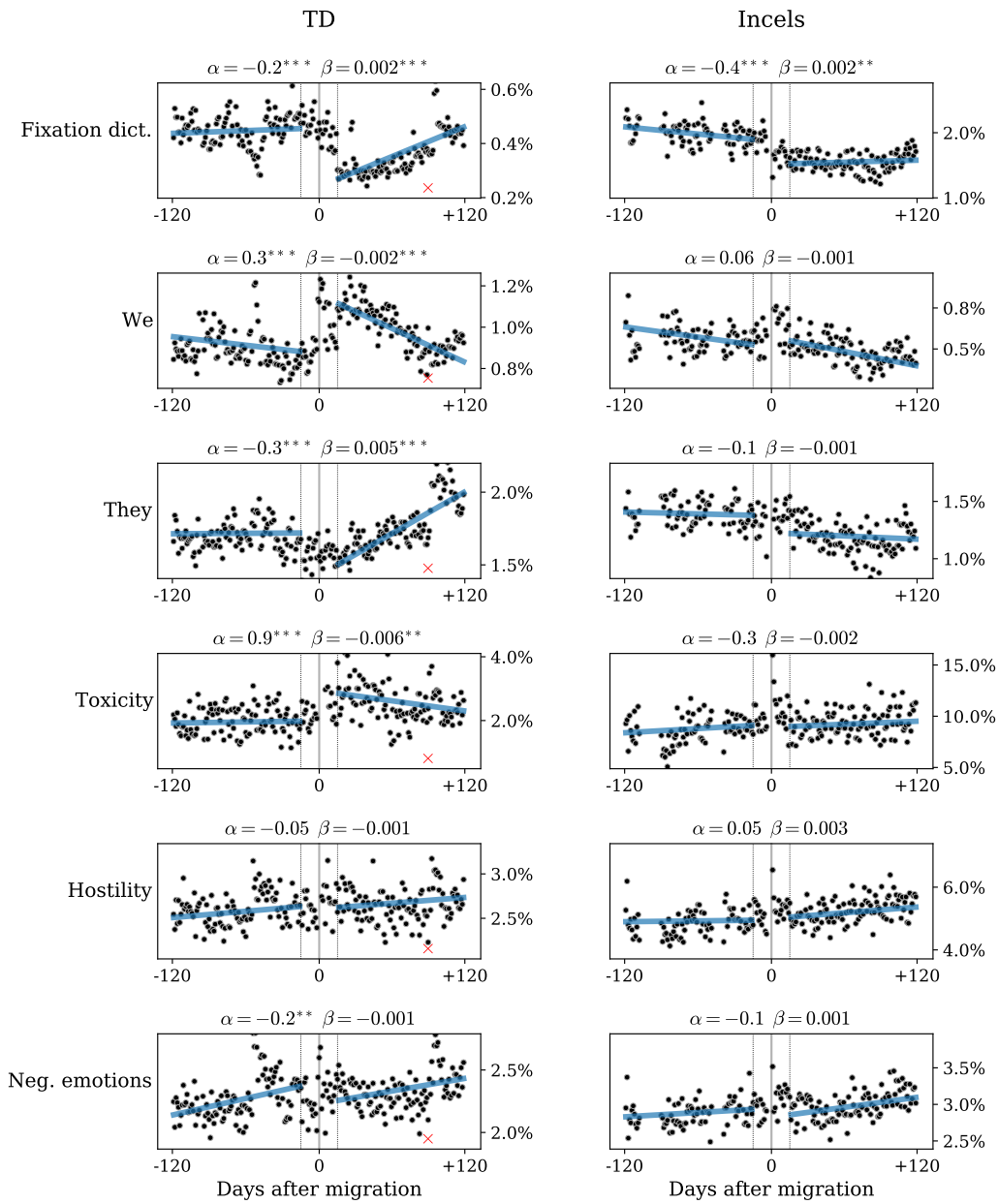

Fig. 6. **Content signals:** Daily content-related statistics for the TD community (left) and the Incel community (right) 120 days before and after migrations. For the *Fixation Dictionary* and the LIWC-related metrics, blue dots depict, for each day, the percentage of words belonging to each word category. For *Toxicity*, they depict the daily percentage of posts with toxicity scores higher than 0.80. For *Toxicity*, *Negative Emotions*, and *Hostility*, we limit our analysis to posts that contain at least one word in our fixation dictionaries. We again show the output of our model as solid blue lines, and the coefficients related to the moderation measure (α and β) on top of each plot (marking those for which $p < 0.001$, 0.01, and 0.05 with ***, **, and *, respectively). For the TD community, we mark the killing of George Floyd, with a red cross (×) close to the x-axis.

5.1 Community-level trends

To study community-level trends, we use a regression discontinuity design similar to Equation (1); however, we add an extra term to control for changes in length associated with the migration.[5] The model now takes on this form:

$$y_t = \alpha_0 + \beta_0 t + \alpha i_t + \beta i_t t + \gamma l_t, \tag{2}$$

where l_t represents the median length of posts (in characters) on day t. We add this covariate to ensure that changes in the intercept (α) and the slope (β) following the intervention are not confounded by changes in the way people post on the new platforms (e.g., longer posts). Note that a consequence of this added term is that when we plot the number of posts (on the y-axis) per day (on the x-axis), we no longer get a straight line since changes in the median length (which varies with time) may impact the outcome of the regression. Thus, for the plots, we fix the value of the length as the average value through the entire period in order to isolate the effect of the intervention and simulate that there is no length change. For descriptions of the other coefficients, see Equation (1). Again, all coefficients for the regression analysis, along with confidence intervals, are shown in Table 3.

Fixation dictionary. We begin by inspecting the prevalence of the fixation dictionary terms over time, as depicted in the first column of Fig. 6. For the TD community we observe a significant drop of $\alpha = -0.2$ percentage points in the usage of terms in the fixation dictionary (-44% of the *MVBI*). For the Incel community, following the intervention, we also see a decrease of around $\alpha = -0.4$ percentage points in the usage of words in the fixation dictionary (-21% of the *MVBI*). In both cases, we observe a positive increase in the trend after the intervention ($\beta = 0.002$ for both communities).

Fixation-related signals. Next, we study changes in *Toxicity*, *Negative Emotions*, and *Hostility*. We limit this analysis to the set of posts containing at least one word in the fixation dictionary (see Table 2) since we are particularly interested in how the communities are talking about their objects of fixation. We consider a comment to be toxic if it has a toxicity score above 80% and calculate, for each day, the fraction of toxic posts. This threshold has been used as a default in other papers [56] and production-ready applications that use the API [37]. For the other LIWC-based metrics, we calculate the proportion of words in the specific dictionaries used per day.

The second column in Fig. 6 shows the changes in the percentage of toxic posts for both communities. For the Incel community, we find no significant change following the interventions. For TD, there is a significant increase right after the intervention of around $\alpha = 0.9$ more toxic posts containing the fixation dictionary (42% of the *MVBI*). However, we see a significant decreasing trend of around $\beta = -0.006$ fewer toxic posts containing words in the fixation dictionary per day. This decrease in the overall trend does not necessarily mean that the average percentage of toxic posts will return to the pre-migration levels. After the sharp increase in toxicity following the moderation measure, the daily toxicity levels may settle at a new baseline higher than pre-migration values.

The third and fourth columns in Fig. 6 depict changes in Negative Emotions and Hostility, respectively. We find that in most cases these two metrics experience a *decrease* in the intercept following community level interventions, although effects are not always significant ($p > 0.05$).

Pronoun usage. In the fifth and sixth columns of Fig. 6, we report the usage of two types of personal pronouns: first-person plural pronouns (e.g., "we," "us," "our") and third-person plural pronouns (e.g., "they," "their"). For the Incel community, we see no significant change in the usage

[5] We find significant changes in the average post length pre- vs. post-migration: 131.7 vs. 118.0 for Incels, and 129.2 vs. 141.2 for TD.

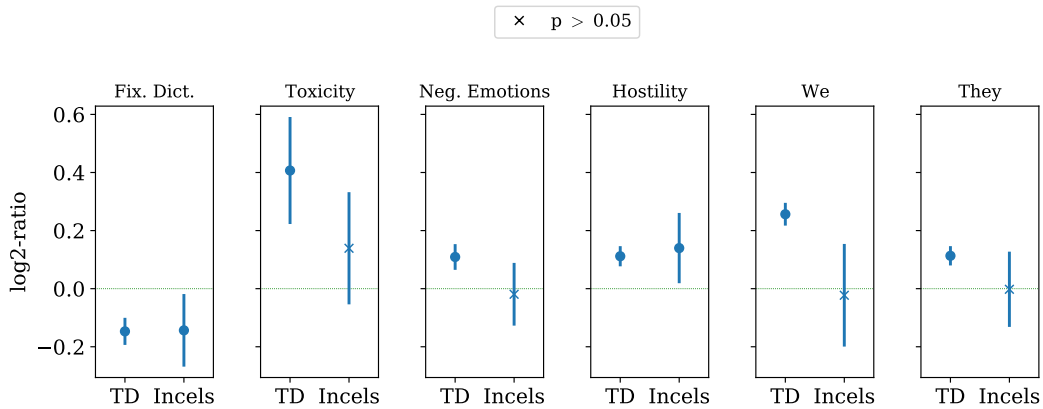

Fig. 7. **User-level change in content:** We depict the mean user-level log-ratio for each of the content-related signals studied. A green horizontal line depicts the scenario of no change (log-ratio = 0). Error bars represent 95% CIs.

of either type of pronoun following the migration. For TD, however, there are interesting changes in their usage. For first person plural pronouns, following the intervention, we find a significant increase in usage of around $\alpha = 0.3$ percentage points (33% of the *MVBI*), and a significant *decrease* in the slope, $\beta = -0.002$. For third-person plural pronouns, we find the opposite. Following the intervention, we find a significant *decrease* of $\alpha = -0.3$ percentage points (−18% of the *MVBI*), followed by a significant increase in the trend, $\beta = 0.005$.

First-person plural pronouns capture group identification and third-person plural pronouns have been associated with extremism [11, 18, 35]. Thus, for the TD community, the intervention seems to have transiently increased group identification immediately after the ban, and later, attention seems to have shifted to the outgroup. The reduced focus on the out-group following the community intervention could also be related to the way words in the *fixation dictionary* were used after migration. There too, we observe a similar pattern: a sharp drop followed by a gradual increase in usage.

Overall, these findings suggest that the community migrations heterogeneously impacted the communities at hand. While not much changed for the Incel community, we find that for TD, there were significant increases in signals related to both the fixation warning behavior (*Toxicity*) and the group behavior warning behavior (both first- and third-person plural pronouns). Again, here a potential confound is the death of George Floyd on 25 May 2020, which impacted user activity (see Fig. 3) and coincides with increases in some of the metrics studied (e.g., third person plural pronouns). By repeating the analysis for TD excluding the period after 24 May 2020, we still find that these changes hold.

5.2 User-level trends

Similar to our content-level analysis, the reasons behind the increase in some of the signals related to online radicalization are important. Here, again, it could be that the subset of users who migrated to the fringe platform were more radical to begin with *or* that the users became more radical after the migration. Thus, it is important to analyze changes at the user level. Luckily, the sample of matched users gives us the opportunity to control for self-selection since we can measure, e.g., the percentage of toxic posts before vs. after the migration for the same group of matched users.

Matched comparison. To disentangle self-selection from user-level increases following the migration, we compare changes in each of the signals for the set of matched users. We calculate, for each user, the fraction of toxic posts (*Toxicity* higher than 0.8), and the percentage of words used in each of the defined categories (*Hostility*, *We*, etc.) both before and after the migration. Then, similar to Fig. 5, we compare the log-ratio between the signals associated with each user *before* and *after* the migration. However, here, calculating the log-ratio may involve dividing by 0, e.g., for a user who posted no toxic posts before the migration and 2 after. Thus, for each individual signal, we limit our analysis to users with positive values for that signal before and after the migration. Therefore, when comparing the changes in toxic posts, we consider only users with at least one toxic post before and one toxic post after the migration. Similarly, for the LIWC-related signals, we consider only users who used words in the given category at least once before and at least once after the migration. We report the mean log-ratio across matched users for each signal in Fig. 7.

For the TD community, we again observe significant increases for *Toxicity* ($\mu = 0.41$, which represents an increase of around 32% since $2^{0.41} \approx 1.32$), *We* ($\mu = 0.26$, 20% increase), and *They* ($\mu = 0.11$, 8% increase). This suggests that the increases previously observed were not caused merely by self-selection. For the Incel community, there were non-significant increases in *Toxicity* (0.13, 9% increase) and small non-significant decreases in the usage of both pronoun-related categories. For both communities, we again significant decreases in the usage of words in the fixation dictionary ($\mu = 0.15$ for TD and $\mu = 0.14$ for Incels, around 11% increase in both cases). We also find significant increases in signals that we did not observe in the community-level analysis. Namely, for both communities we find significant increases in *Hostility* (8% increase for TD and 10% for Incels), and for TD, we find a significant increase in *Negative Emotions* (8% increase).

Regression discontinuity analysis. The previous analysis indicates that there were significant changes in the radicalization-related signals for the matched sample, some of which we did not observe in the community-level analysis. To better understand the matched sample, and the differences between the results at the community-level and the user-level, we repeat the regression discontinuity analysis done for the signals of interest using only posts from the matched user sample. We use exactly the same model as in Equation (2), changing only the data: *there* we used all posts by all users, *here* we use all posts by matched users. In Fig. 8, we plot the regression lines for the analysis done with all users in blue and for matched users in orange. Coefficients along with confidence intervals are again presented in Table 3.

For several signals, the results in this reduced sample are very similar to the previous analysis. For example, for TD, we have almost exactly the same coefficients for the usage of the fixation dictionary ($\alpha = -0.2$, and $\beta = 0.002$) and of third-person plural pronouns ($\alpha = -0.3$, and $\beta = 0.004$). Yet, for some of the signals, we do find significant differences following the community migrations. More specifically, for TD community, following the migration, we find significant increases in the trends for Negative Emotions ($\beta = 0.002$) and we find no significant *decrease* in the trend for the Toxicity signal (which used to be the case). Also, for Incels, we find significant increases in the trend for *Negative Emotions* ($\beta = 0.004$) and *Hostility* ($\beta = 0.01$).

Overall, this analysis confirms the results previously discussed in Fig. 7 and suggest that users in the matched sample were disproportionally impacted by the community-level intervention. This is different from what we observed when looking at activity levels. There, when we zoomed in on matched users, we found that they had *decreased* their activity (even though the number of posts per user grew). Here, on the contrary, we find that these users seem to have become more radical than the general population.

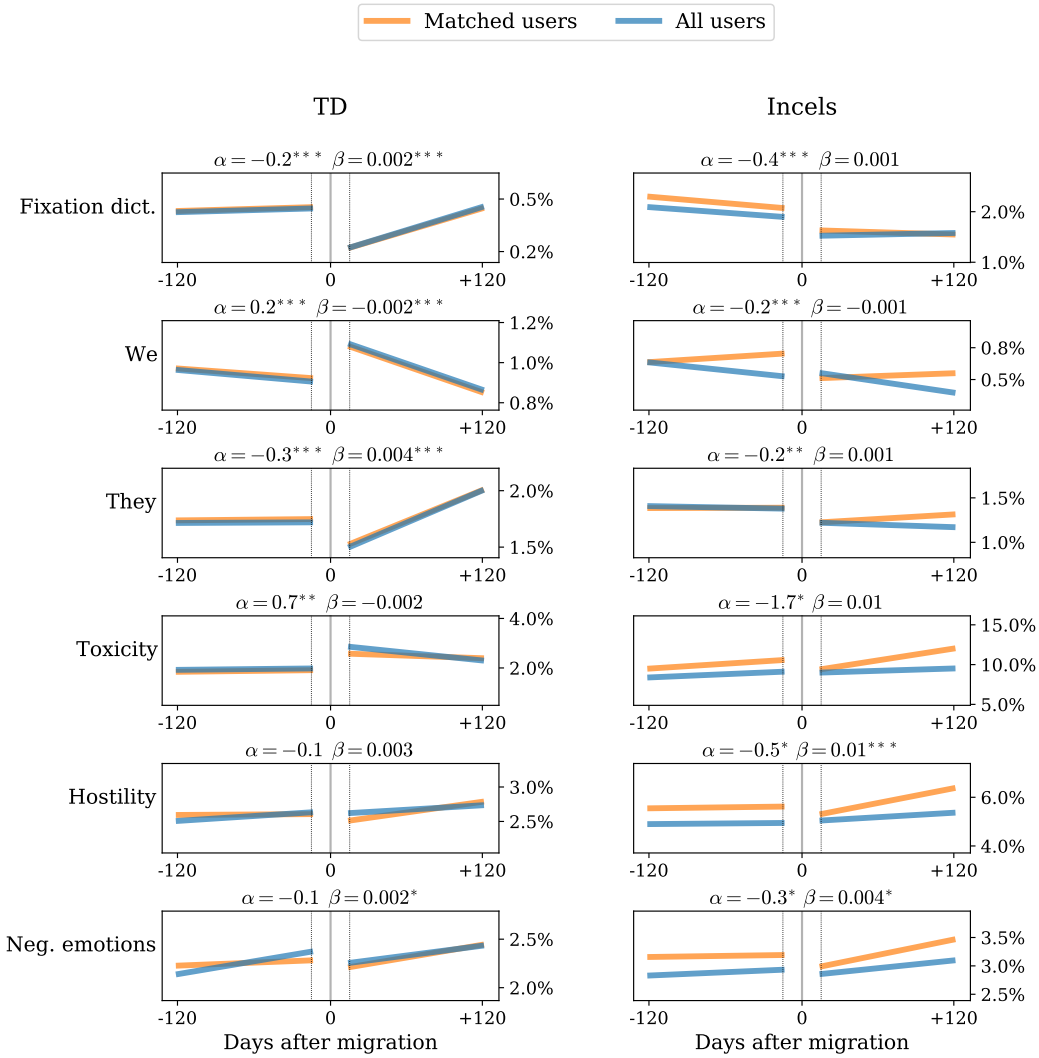

Fig. 8. **Daily content signals for matched users:** We repeat the same analysis from Sec. 5.1 considering the sample of matched users. We show the regression lines considering all users (in blue) and only matched users (in orange). Above each plot, we show the coefficients related to the moderation (α and β) for the model considering only matched users. For additional details, see Fig. 6.

5.3 Take-aways

Altogether, our analysis shows that, for TD, community-level interventions and the migrations that ensued are associated with significant increases in radicalization-related signals. A closer look at the matched user sample indicates that these increases were not merely due to self-selection, since we also observe significant user-level increases for all but one of the signals studied. Also, analyzing the matched sample, we find that the migration may have disparately impacted these users, since the differences for them are more substantial.

A second important result of our content-level analysis is that communities were heterogeneously impacted. When comparing how the *activity* in the two communities changed (Sec. 4), we found the same patterns overall; whereas, when comparing how the *content* changed, we found rather distinct behaviors across the two communities. Unlike the TD community, for Incels, there were often *decreases* in signals related to radicalization following community migration.

6 DISCUSSION & CONCLUSION

Our work paints a nuanced portrait of the benefits and possible backlashes of community-level interventions. On the one hand, we found that the interventions were effective in decreasing activity and the capacity of the community to attract newcomers. Also, we found evidence that *relative* increase in activity (i.e., fewer users posting more) is likely due to self-selection: the users who migrated to the new community were more active to begin with. On the other hand, we found significant increases in radicalization-related signals for one of the communities studied (TD), even when controlling for self-selection. In fact, these increases were even more substantial for the set of matched users studied.

An interesting angle to consider the changes observed in communities pre- vs. pos- migration, is through the lens of characteristics and affordances of large online platforms such as Reddit, YouTube and Facebook. According to Gillespie [15], online platforms differ from traditional media outlets in that they provide the means of distribution, but not the content (which is user generated). Moreover, a key component of online platforms is that they moderate and gatekeep content, despite best efforts to present themselves as neutral "facilitators."

In that context, the migration of online communities from mainstream platforms to fringe, alternative websites provoke shifts associated with how content is distributed and moderated, two important roles of online platforms. The decrease in activity observed after communities migrated emphasizes the *power of the distribution* of online platforms such as Reddit. Since Reddit has thousands of highly popular subreddits, toxic communities inhabiting the platform are easily discoverable, and consuming the content they produce is convenient. Using a similar line of reasoning, the increase in toxicity observed when members of r/The_Donald migrated out of the subreddit can be associated with the *power of the moderation* of online platforms. Toxicity may be understood as a proxy for content that would likely clash with Reddit's content policy[6]. Therefore, the rise in toxicity following the ban can be understood as a consequence of the removal of platform moderation.

Overall, our results strengthen the hypothesis that platforms are largely responsible for our online information ecosystem [15]. Besides determining what kinds of content flourishes [32], platforms allow communities to exploit their affordances to recruit new members, and are able to influence the content being posted in toxic communities. In the remainder of this section, we discuss the implications of these results for platforms and future research, as well as the limitations of our study.

6.1 Limitations and future work

Communities. Our work focuses on two communities: TD and Incels. However, Reddit has sanctioned many other communities that may have migrated to new fringe websites. The implications of such sanctions for migration may differ based on the specifics of each community. That said, the communities we study are among the most prominently sanctioned subreddits, and our analysis provides early insights into the consequences of such sanctions. In the future, similar analysis on

[6] https://www.reddit.com/r/reddit.com/wiki/revisions/contentpolicy

other sanctioned communities would help disentangle how contextual factors including community size, topic, and the design of the alternative platform may affect migration patterns.

Migrations and dispersion. We consider the effects of migration to only one fringe website per each of the sanctioned communities we study. In both cases, the migrations to the websites we analyze were officially endorsed by the subreddits' moderators, and, for r/The_Donald, the subreddit promoted the migration to the new site while it could. However, users may have migrated to other platforms as well. For example, on Reddit, after r/Incels was banned, an old subreddit called r/Braincels reportedly became popular (until eventually being banned too). Also more broadly, some community-level interventions may not result in "successful" coordinated migration. Rather, users can be dispersed through a variety of other platforms (e.g., Gab, 4chan, Parler, etc.). Studying what happens in these cases is an important direction to completely understand the impact of deplatforming communities. For example, one could try to measure the activity boost experienced in each of these platforms whenever a toxic community in a mainstream platform (e.g. r/The_Donald) gets banned. A challenge here would be to obtain data for a variety of fringe communities and to control for other confounders, such as geopolitical events.

Confounders. The responsiveness of these communities to real-world events creates confounders. This is particularly true for the TD community, where we found significant changes in the content- and activity-related signals in reaction to the killing of George Floyd. While our quasi-experimental research design controls for linear trends, sudden bursts in content-related signals can partially impact our results. Controlling for these trends is hard since the reaction of these communities to real world changes is inherently linked to the harms they pose to society. However, in our specific case, we find that the effects observed held even when limiting the period of the regression discontinuity analysis to before the event (i.e., George Floyd's killing). Another possible set of confounders are changes to rules and moderation actions that could have changed pre- vs. post-intervention. Although we did not explicitly incorporate these changes into our analysis, we carefully analyzed the set of rules before (Incels: [43], TD: [46]) and after (Incels: [22], TD: [53]) the migration and did not find any substantial changes.

Matched Users. Another limitation of the work at hand is that user-level analyses are made on a set of users matched according to their usernames. These users tend to be more active than the average user (cf. Fig. 4), and may differ from users who migrated and did not change their username. Although important, we argue that this bias is not impactful to the external validity of our results. The main purpose of looking at matched users is to distinguish between behavior change and self-selection. When studying changes in activity, analyzing matched users provides us with the useful insight that, although the number of posts per user increases after migration (cf. Fig. 3), on a user-level this is not the case (cf. Fig. 5). For the sample bias to be an issue here, reverting or weakening the results, it would be necessary that users who migrated and did not keep the same username became more active after the ban while those who kept the same username did not, which is unlikely. When studying changes in the content, we find that community-level trends on the set of matched users are very similar to community-level changes considering all users (cf. Fig. 8), weakening concerns that there would be a strong difference between the nature of the content posted by these users. An interesting direction to further understand these matched users (and explore user-level trends) would be to additionally analyze users with known usernames pre- vs. post- ban in other subreddits.

Mapping signals to externalities. Our analysis relies on user activity and signals derived from user-generated content to analyze online toxic communities. Our main result suggests that community-level interventions may involve a trade-off: less activity at the expense of a more radical

community elsewhere. Yet, the relationship between these activity- and content-related signals from toxic online communities and their real-world harms is still fuzzy. It is unclear, for instance, whether a reduction of 50% in posting activity where each user is 10% more "toxic" is desirable or not. While such a fine-grained assessment of the consequences of a moderation intervention is out of the scope of this paper, further study of the causal links between toxicity, user activity, and real-world harm is an important research direction to improve the quality of moderation decisions.

6.2 Implications for online platforms

Our analysis of migration dynamics highlights that community-wide moderation interventions do not happen in a vacuum. When platforms sanction an entire community, as opposed to taking user-level actions, communities may migrate *en gros* to a different platform. Platforms have difficult decisions to make: they need to consider the effects of community-wide sanctions not only on their own backyard, but on other online and offline spaces as well. Our results suggest that there may be a trade-off associated with this decision: banning a community from a mainstream platform may come at the expense of a smaller but more extreme community elsewhere. However, this take-away should be handled with nuance, since our work is limited to two communities, and since the increase in toxicity was only observed in one of the two communities.

Nevertheless, a practical implication that follows from our results is that, given that a community eventually gets banned, the time the said community was allowed to flourish in a mainstream platform may increase its potential for harm post-banning. The reasoning is simple: since community growth is halted by the deplatforming, the earlier the community is banned, the fewer members a possible spin-off community would have. In that context, if banning is a commonly used practice in a given platform, it is advantageous to employ the measure proactively rather than reactively.

Lastly, the methodological framework we use in this paper may also be used in other contexts and platforms to evaluate the effectiveness of moderation interventions. Platforms have at their disposal abundant data that can help further clarify the trade-offs we discussed here. We hope that extensions of this work will yield more precise guidelines on how to handle problematic online communities.

ACKNOWLEDGEMENTS

Manoel Horta Ribeiro is supported by a Facebook Fellowship Award. Jeremy Blackburn is supported by NSF grants CNS-2114411 and IIS-2046590. Gianluca Stringhini is supported by NSF grants CNS-1942610 and CNS-2114407. Emiliano De Cristofaro is supported by the UK's National Research Centre on Privacy, Harm Reduction, and Adversarial Influence Online (REPHRAIN, UKRI grant: EP/V011189/1). Robert West is partly supported by a grant from the EPFL/UNIL Collaborative Research on Science and Society (CROSS) Program, the Swiss National Science Foundation (grant 200021_185043), the European Union (TAILOR, grant 952215), and gifts from Google, Facebook, and Microsoft.

REFERENCES

[1] Almeida, V., Filgueiras, F., and Gaetani, F. Digital governance and the tragedy of the commons. *IEEE Internet Computing* (2020).

[2] Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The Pushshift Reddit dataset. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2020).

[3] Blackwell, L., Dimond, J. P., Schoenebeck, S., and Lampe, C. Classification and its consequences for online harassment: design insights from HeartMob. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2017).

[4] Borah, P. Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research* (2014).

Table 3. Coefficients for all regression discontinuity analyses done throughout the paper, including 95% confidence intervals. Coefficients for which $p < 0.001$, 0.01, and 0.05 are marked with ***, **, and *, respectively. The value $[10^{-3}]$ in the beginning of a cell indicates that the value of the cell as well as the confidence intervals presented should be multiplied by 10^{-3}. This may cause slight differences in the numbers in this table and the ones presented in the plots, since here we present the results at higher precision. Note that this table contains the regression results for three different analysis carried out throughout the paper and depicted in Fig. 3, Fig. 6, and Fig. 8. For presentation reasons we omit the confidence intervals for the intercept across the whole period (α_0), which is significant ($p < 0.001$) across all of the models.

(a) Community-level activity (Fig. 3)

Venue	Statistic	α_0	β_0	α	β	R^2
TD	#newcomers	214.7^{***}	$-0.5\,(-1.3, 0.2)$	$-77.8^{**}\,(-127.9, -27.7)$	$0.5\,(-0.5, 1.4)$	0.24
	#posts	26650^{***}	$8.8\,(-17.9, 35.5)$	$-14416^{***}\,(-16947, -11886)$	$120.6^{***}\,(83.9, 157.2)$	0.41
	#users	7593^{***}	$4\,(-0.5, 8.5)$	$-4774^{***}\,(-5184, -4365)$	$13.7^{***}\,(8.1, 19.3)$	0.85
	#posts/#users	3.5^{***}	$-0.001\,(-0.003, 0.001)$	$1.1^{***}\,(0.9, 1.3)$	$0.009^{***}\,(0.006, 0.01)$	0.85
Incels	#newcomers	224.9^{***}	$0.9^{***}\,(0.5, 1.4)$	$-215.4^{***}\,(-250.3, -180.5)$	$-0.9^{***}\,(-1.3, -0.4)$	0.84
	#posts	4840^{***}	$19.5^{***}\,(14.1, 25)$	$-2651^{***}\,(-3098, -2203)$	$1.7\,(-4.8, 8.2)$	0.55
	#users	960.1^{***}	$3.9^{***}\,(2.9, 4.9)$	$-777.4^{***}\,(-850.2, -704.6)$	$-3^{***}\,(-4, -2.1)$	0.93
	#posts/#users	5^{***}	$-0.001\,(-0.003, 0.003)$	$7.4^{***}\,(6.6, 8.3)$	$0.04^{***}\,(0.02, 0.05)$	0.92

(b) Community-level content (Fig. 6)

Venue	Statistic	α_0	β_0	α	β	R^2
TD	Fix. Dict	0.6^{***}	$[10^{-3}]0.2\,(-0.2, 0.5)$	$-0.2^{***}\,(-0.3, -0.2)$	$[10^{-3}]1.7^{***}\,(1.2, 2.2)$	0.53
	Toxicity	3.1^{***}	$[10^{-3}]0.5\,(-2, 3.1)$	$0.9^{***}\,(0.6, 1.3)$	$[10^{-3}]-5.9^{**}\,(-10.1, -1.7)$	0.3
	Neg. Emotion	2.7^{***}	$[10^{-3}]2.2^{***}\,(1.1, 3.3)$	$-0.2^{**}\,(-0.3, -0.06)$	$[10^{-3}]-0.5\,(-2.1, 1.1)$	0.14
	Hostility	3.3^{***}	$[10^{-3}]1.2\,(-0.04, 2.4)$	$-0.05\,(-0.2, 0.09)$	$[10^{-3}]-0.1\,(-2, 1.8)$	0.08
	We	0.6^{***}	$[10^{-3}]-0.7^{**}\,(-1.2, -0.2)$	$0.3^{***}\,(0.2, 0.3)$	$[10^{-3}]-2.1^{***}\,(-2.7, -1.4)$	0.46
	They	1.3^{***}	$[10^{-3}]0.05\,(-0.5, 0.6)$	$-0.3^{***}\,(-0.4, -0.2)$	$[10^{-3}]4.8^{***}\,(3.8, 5.7)$	0.55
Incels	Fix. Dict	1.9^{***}	$[10^{-3}]-1.8^{*}\,(-3.3, -0.4)$	$-0.4^{***}\,(-0.5, -0.2)$	$[10^{-3}]2.4^{**}\,(0.9, 3.8)$	0.69
	Toxicity	11.4^{***}	$[10^{-3}]6.9\,(-3.1, 16.8)$	$-0.3\,(-1.2, 0.6)$	$[10^{-3}]-1.9\,(-13.7, 9.9)$	0.13
	Neg. Emotion	3.2^{***}	$[10^{-3}]1\,(-0.7, 2.7)$	$-0.1\,(-0.3, 0.03)$	$[10^{-3}]1.3\,(-0.4, 3.1)$	0.25
	Hostility	6.3^{***}	$[10^{-3}]0.4\,(-3.1, 4)$	$0.05\,(-0.3, 0.4)$	$[10^{-3}]2.6\,(-0.8, 6.1)$	0.38
	We	0.6^{***}	$[10^{-3}]-1\,(-2.5, 0.4)$	$0.06\,(-0.05, 0.2)$	$[10^{-3}]-0.4\,(-1.3, 0.5)$	0.3
	They	1.2^{***}	$[10^{-3}]-0.3\,(-2.6, 2.1)$	$-0.1\,(-0.3, 0.04)$	$[10^{-3}]-0.2\,(-1.6, 1.2)$	0.44

(c) Community-level content for matched users (Fig. 8)

Venue	Statistic	α_0	β_0	α	β	R^2
TD	Fix. Dict	$0.6^{***}\,(0.5, 0.8)$	$[10^{-3}]0.2\,(-0.2, 0.5)$	$-0.2^{***}\,(-0.3, -0.2)$	$[10^{-3}]1.6^{***}\,(1.1, 2.1)$	0.54
	Toxicity	$3.2^{***}\,(2.1, 4.3)$	$[10^{-3}]0.6\,(-2.8, 4)$	$0.7^{**}\,(0.2, 1.2)$	$[10^{-3}]-2.4\,(-9, 4.3)$	0.13
	Neg. Emotion	$2.5^{***}\,(2.2, 2.8)$	$[10^{-3}]0.5\,(-0.5, 1.6)$	$-0.1\,(-0.2, 0.006)$	$[10^{-3}]1.7^{*}\,(0.06, 3.3)$	0.08
	Hostility	$3.1^{***}\,(2.7, 3.5)$	$[10^{-3}]0.1\,(-1.4, 1.6)$	$-0.1\,(-0.3, 0.04)$	$[10^{-3}]2.5\,(-0.001, 5)$	0.06
	We	$0.5^{***}\,(0.2, 0.8)$	$[10^{-3}]-0.6\,(-1.2, 0.007)$	$0.2^{***}\,(0.2, 0.3)$	$[10^{-3}]-2.2^{***}\,(-2.9, -1.4)$	0.37
	They	$1.6^{***}\,(1.3, 1.8)$	$[10^{-3}]0.1\,(-0.6, 0.8)$	$-0.3^{***}\,(-0.4, -0.2)$	$[10^{-3}]4.5^{***}\,(3.4, 5.6)$	0.44
Incels	Fix. Dict	$1.9^{***}\,(1.6, 2.2)$	$[10^{-3}]-2.1^{*}\,(-3.9, -0.4)$	$-0.4^{***}\,(-0.6, -0.2)$	$[10^{-3}]1.4\,(-0.9, 3.6)$	0.68
	Toxicity	$11^{***}\,(7.9, 14.1)$	$0.01\,(-0.008, 0.03)$	$-1.7^{*}\,(-3.2, -0.1)$	$0.01\,(-0.007, 0.04)$	0.08
	Neg. Emotion	$3^{***}\,(2.6, 3.4)$	$[10^{-3}]0.3\,(-3, 3.7)$	$-0.3^{*}\,(-0.5, -0.04)$	$[10^{-3}]4.2^{*}\,(0.6, 7.8)$	0.1
	Hostility	$6^{***}\,(5.3, 6.7)$	$[10^{-3}]0.7\,(-3.9, 5.3)$	$-0.5^{*}\,(-0.9, -0.06)$	$[10^{-3}]9.5^{***}\,(4, 15.1)$	0.2
	We	$0.4^{***}\,(0.3, 0.6)$	$[10^{-3}]0.6\,(-0.6, 1.8)$	$-0.2^{***}\,(-0.3, -0.1)$	$[10^{-3}]-0.3\,(-1.5, 1)$	0.44
	They	$0.9^{***}\,(0.7, 1.2)$	$[10^{-3}]0.04\,(-1.7, 1.8)$	$-0.2^{**}\,(-0.3, -0.04)$	$[10^{-3}]0.8\,(-1.2, 2.8)$	0.28

[5] Brewster, J. Forbes — The extremists, conspiracy theorists, and conservative stars banned from social media following the capitol takeover. https://bit.ly/3xPvSLJ, 2020.

[6] Carillo, K. D. A., and Marsan, J. The dose makes the poison: exploring the toxicity phenomenon in online communities. In *ICIS* (2016).

[7] Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. You can't stay here: the efficacy of Reddit's 2015 ban examined through hate speech. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2017).

[8] Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., and Leskovec, J. Anyone can become a troll: causes of trolling behavior in online discussions. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2017).

[9] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. Antisocial behavior in online discussion communities. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2015).

[10] Clegg, N. Facebook Newsroom — Welcoming the Oversight Board. https://bit.ly/2VVgHU7, 2020.

[11] Cohen, K., Johansson, F., Kaati, L., and Mork, J. C. Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence* (2014).

[12] Davidson, T., Warmsley, D., Macy, M., and Weber, I. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2017).

[13] Dewey, C. Washington Post — These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy, and why it banned them. https://wapo.st/3AO7pbl, 2016.

[14] Flores-Saviaga, C. I., Keegan, B. C., and Savage, S. Mobilizing the Trump Train: understanding collective action in a political trolling community. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2018).

[15] Gillespie, T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.* Yale University Press, 2018.

[16] Gilmour, D. The Daily Dot — 4chan and Reddit users set out to prove Seth Rich murder conspiracy. https://bit.ly/2VWJ4B8, 2017.

[17] Ging, D. Alphas, betas, and Incels: Theorizing the masculinities of the Manosphere. *Men and Masculinities* (2019).

[18] Grover, T., and Mark, G. Detecting potential warning behaviors of ideological radicalization in an Alt-Right subreddit. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2019).

[19] Hatebase. Hatebase. https://www.hatebase.org/, 2018.

[20] Hines, A. The Cut — How many bones would you break to get laid? https://bit.ly/2VSTrpD, 2019.

[21] Hoffman, B., Ware, J., and Shapiro, E. Assessing the threat of Incel violence. *Studies in Conflict & Terrorism* (2020).

[22] incels.co. Rules. https://bit.ly/3yY8wF2, 2018.

[23] Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. Online harassment and content moderation: the case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* (2018).

[24] Johnson, N. F., Leahy, R., Restrepo, N. J., Velasquez, N., Zheng, M., Manrique, P., Devkota, P., and Wuchty, S. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature* (2019).

[25] Kates, N. Statusmaxxing Admincel. https://open.spotify.com/episode/0MsycI8DgqpcX2NJEouTWl, 2019.

[26] Kwak, H., Blackburn, J., and Han, S. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015).

[27] Lewis, R. Alternative influence: broadcasting the reactionary right on YouTube. *Data & Society* (2018).

[28] Lilly, M. *'The World is Not a Safe Place for Men': The Representational Politics Of The Manosphere.* PhD thesis, Université d'Ottawa/University of Ottawa, 2016.

[29] Lyons, M. N. Ctrl-alt-delete: the origins and ideology of the alternative right. *Political Research Associates* (2017).

[30] Massanari, A. #Gamergate and The Fappening: how Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* (2017).

[31] Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., and Mukherjee, A. Hate begets hate: a temporal study of hate speech. In *Proceedings of the ACM on Human-Computer Interaction (CSCW)* (2020).

[32] Munger, K., and Phillips, J. Right-wing YouTube: a supply and demand perspective. *The International Journal of Press/Politics* (2020).

[33] Newell, E., Jurgens, D., Saleem, H. M., Vala, H., Sassine, J., Armstrong, C., and Ruths, D. User migration in online social networks: a case study on Reddit during a period of community unrest. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2016).

[34] Ohlheiser, A. The Washington Post — Fearing yet another witch hunt, reddit bans pizzagate. https://wapo.st/2Xvbvae, 2016.

[35] Pennebaker, J. W., and Chung, C. K. Computerized text analysis of Al-Qaeda transcripts. *A content analysis reader* (2007).

[36] Perspective API. https://www.perspectiveapi.com/, 2018.

[37] Project, C. Media Coral open source commenting platform. https://docs.coralproject.net/talk/toxic-comments/.

[38] Rajadesingan, A., Resnick, P., and Budak, C. Quick, community-specific learning: how distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2020).
[39] Raman, N., Cao, M., Tsvetkov, Y., Kästner, C., and Vasilescu, B. Stress and burnout in open source: toward finding, understanding, and mitigating unhealthy interactions. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results* (2020).
[40] Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., and Zannettou, S. The evolution of the Manosphere across the Web. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2021).
[41] Ribeiro, M. H., Calais, P. H., Santos, Y. A., Almeida, V. A., and Meira Jr, W. Characterizing and detecting hateful users on Twitter. In *Proceedings of the International Conference on Web and Social Media (ICWSM)* (2018).
[42] Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., and Meira Jr, W. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).
[43] r/Incels. Rules. https://bit.ly/3iXIA77, 2018.
[44] Roberts, S. T. *Behind the screen: content moderation in the shadows of social media.* Yale University Press, 2019.
[45] r/OutOfTheLoop. Post: "What's up with /r/The_Donald leaving Reddit?". https://www.reddit.com/r/OutOfTheLoop/comments/6bzv8v/, 2017.
[46] r/The_Donald. Rules. https://bit.ly/3k47MIp, 2018.
[47] r/The_Donald. Post: "Bookmark this site". https://bit.ly/30brHfm, 2020.
[48] Saleem, H. M., and Ruths, D. The aftermath of disbanding an online hateful community. *arXiv:1804.07354* (2018).
[49] Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019).
[50] Shen, Q., and Rose, C. The discourse of online content moderation: investigating polarized user responses to changes in Reddit's quarantine policy. In *Proceedings of the Third Workshop on Abusive Language Online* (2019).
[51] Solon, O. The Guardian — Reddit bans misogynist men's group blaming women for their celibacy. https://bit.ly/2W2M0fq, 2017.
[52] Tausczik, Y. R., and Pennebaker, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* (2010).
[53] thedonald.win. Rules. https://bit.ly/3g7IKH9, 2018.
[54] thedonald.win. Post: "I hope if you came from T_D you reserved your reddit username even if you don't plan to use it". thedonald.win/p/FMA6trrU/, 2020.
[55] Timberg, C., and Dwoskin, E. Washington Post — Reddit closes long-running forum supporting President Trump after years of policy violations. https://wapo.st/3ySp2Gv, 2020.
[56] Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., and Stringhini, G. Measuring and characterizing hate speech on news websites. In *ACM Conference on Web Science* (2020).
[57] Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., and Thain, N. Conversations gone awry: detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (2018).