Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald

AMAURY TRUJILLO, Institute for Informatics and Telematics, National Research Council (IIT-CNR), Italy
STEFANO CRESCI, Institute for Informatics and Telematics, National Research Council (IIT-CNR), Italy

The subreddit `r/The_Donald` was repeatedly denounced as a toxic and misbehaving online community, reasons for which it faced a sequence of moderation interventions by Reddit administrators. It was quarantined in June 2019, restricted in February 2020, and finally banned in June 2020, but despite precursory work on the matter, the effects of this sequence of interventions are still unclear. In this work, we follow a multidimensional causal inference approach, with data containing more than 15M posts made in a time frame of 2 years, to examine the effects of such interventions inside and outside of the subreddit. We find that the interventions greatly reduced the activity of problematic users. However, the interventions also caused an increase in toxicity and led users to share more polarized and less factual news. In addition, the restriction had stronger effects than the quarantine, and core users of `r/The_Donald` suffered stronger effects than the rest of users. Overall, our results provide evidence that the interventions had mixed effects and paint a nuanced picture of the consequences of community-level moderation strategies. We conclude by reflecting on the challenges of policing online platforms and on the implications for the design and deployment of moderation interventions.

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

Additional Key Words and Phrases: content moderation; moderation interventions; online communities; toxicity; news quality; causal inference

ACM Reference Format:
Amaury Trujillo and Stefano Cresci. 2022. Make Reddit Great Again: Assessing Community Effects of Moderation Interventions on r/The_Donald. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 526 (November 2022), 28 pages. https://doi.org/10.1145/3555639

1 INTRODUCTION

The social media and news aggregator platform Reddit is among the most popular Internet websites, ranking as the sixth most visited website and the third most visited social media in the United States, as of April 2022.[1] The platform is organized in communities called subreddits, in which users submit and discuss content regarding the community's shared topics and interests. Subreddits cover nearly every aspect of life, including news, sports, science, technology, religion, and a broad spectrum of social and other activities. Overall, *politics* is one of the most discussed topics on the platform, with several subreddits related to US politics consistently ranking among the most popular communities on the platform. The outreach of these communities is enormous, reaching millions

[1] https://www.alexa.com/topsites/countries/US

Authors' addresses: Amaury Trujillo, Institute for Informatics and Telematics, National Research Council (IIT-CNR), via G. Moruzzi 1, Pisa, Italy, 56124, amaury.trujillo@iit.cnr.it; Stefano Cresci, stefano.cresci@iit.cnr.it, Institute for Informatics and Telematics, National Research Council (IIT-CNR), via G. Moruzzi 1, Pisa, Italy, 56124.

2573-0142/2022/11-ART526 $15.00
https://doi.org/10.1145/3555639

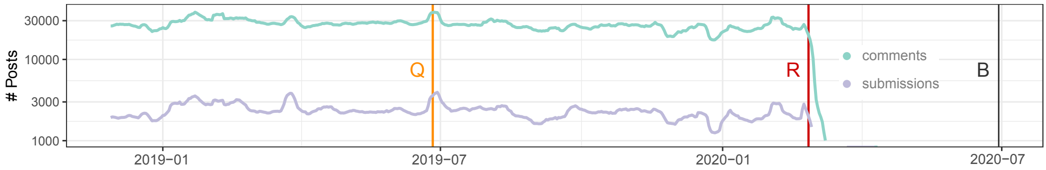


Fig. 1. Seven-day running average of daily posts in r/The_Donald around quarantine (Q), restriction (R), and ban (B). The number of submissions plummeted after the restriction, with comments following suit a few days later. Upon banning, the subreddit had already been completely inactive for several weeks.

of users on Reddit, other online platforms (e.g., Facebook, Twitter) [55], and even the audience of traditional media [49]. For these reasons, political subreddits such as r/politics, r/The_Donald, r/conservative, and others, have received much scholarly attention [28, 41, 49].

In addition to platform-wise rules and policies, each subreddit sets its own behavior guidelines. Furthermore, unlike other platforms, users on Reddit can create and moderate their own subreddits in collaboration with others. As such, each community presents unique characteristics and develops its own habits, participation culture, and moderation rules [49]. Occasionally, some communities accept and even encourage aggressive and harmful behaviors. When such communities repeatedly violate Reddit's policies, platform administrators (i.e., Reddit personnel) intervene to moderate the subreddit. Due to the sensitive and politicized nature of such behaviors, some readers might find upsetting some phrases used herein for their characterization.

1.1 Moderating r/The_Donald

The community of Donald Trump supporters of r/The_Donald was repeatedly denounced for toxicity, trolling, and harassment [22, 28, 34, 36]. For such misbehavior, Reddit administrators imposed a sequence of increasingly restrictive moderation interventions, as shown in Figure 1:

(1) *Quarantine (June 26, 2019)*: r/The_Donald was quarantined following repeated reports for inciting violence, including threatening US public figures. The subreddit was removed from the platform's search results and from the feed of non-subscribed users, albeit these could still access its content if they opted-in after receiving a warning message upon visit [10]. These measures intended to reduce the subreddit's visibility and to deter newcomers.
(2) *Restriction (February 26, 2020)*: Administrators further restricted the subreddit and removed several of its moderators who were supporting content in violation of Reddit's policies, allowing new submissions to be posted only by approved users. Participation in the subreddit came to a complete halt within the following weeks, with the majority of users migrating to other subreddits or even to completely different platforms [28]. The existing content remained accessible for reading and commenting after the restriction.
(3) *Ban (June 29, 2020)*: r/The_Donald was banned, together with other two thousand subreddits, as part of Reddit's actions to enforce new policies. This ban permanently shut the subreddit down, by removing it from the platform and making it impossible to access its contents.

In addition to raising ethical and legal concerns as to whether online platforms should be allowed to limit the freedom of speech of their users, including that of major politicians [8, 26], the practical consequences of these interventions are still unclear. Obviously, platforms apply moderation interventions to reduce the spread of toxic, hateful, fake, and otherwise problematic content [29]. However, the extent to which such interventions are capable of mitigating the issues is still up to debate, and even more so given that certain interventions caused opposite effects to those planned —i.e., they *backfired* [2, 9, 21, 40]. For these reasons, a growing body of studies has recently focused on evaluating social media moderation strategies [10, 11, 28, 29, 45].

While these early works provided interesting results, many questions remain unanswered. Firstly, interventions may produce effects spread across multiple dimensions of user behavior and ideology [33]. For instance, they can affect user activity and participation to certain communities, political preferences, news consumption habits, and more. Early literature mainly investigated intervention effects with respect to user *activity* and *toxicity* (including hate speech) [10, 11, 28, 29, 45]. However, many other important dimensions could be considered, such as the degrees of *political polarization* and *factual reporting*, which are known drivers of online misbehavior [2]. Currently, it is unknown whether past interventions produced any effect in these dimensions. Secondly, many different types of interventions are adopted by platform administrators, but the majority of existing studies only analyzed deplatforming interventions [11, 28, 29, 45] —that is, interventions aimed at removing problematic users or communities from a platform, such as Reddit's restrictions and bans. The effects (or lack thereof) of other interventions, such as quarantines, are still unclear [10]. Thirdly, existing studies evaluated each intervention in isolation. However, certain interventions are enforced as part of a sequence of actions, as in the case of `r/The_Donald`. Finally, interventions that do not completely shut a community (e.g., quarantines, as opposed to bans) may produce effects both within and outside the moderated community, because most users browse and interact with multiple communities at the same time. Thus, when a community suffers a moderation intervention, the behavior of the members of that community might change not only within *that* community, but also in *other communities* in which they participate. Regarding Reddit's interventions on `r/The_Donald`, it would be interesting to go beyond existing results by evaluating the intervention effects outside `r/The_Donald`. Overall, additional analyses are needed to improve our understanding of the effects of recent moderation interventions.

1.2 Research Questions

The present study complements the existing literature in some of the aforementioned directions. In particular, Reddit's quarantines are almost completely unexplored, since the majority of existing studies focused on restrictions and bans [11, 28, 29, 45]. We extend and complement the only existing related study [10] by investigating the effects of Reddit's quarantine on `r/The_Donald` with respect to the toxicity of the affected users and by differentiating between different types of users. For this reason, we ask the following starting question:

RQ1: What were the effects of the quarantine, in terms of activity and toxicity, within `r/The_Donald`?

Following from this question, and given that focusing only on a small set of metrics could be misleading, we expand current literature by measuring effects also with respect to the quality of the news articles consumed by members of a community, in order to gain deeper insights into the effects that moderation interventions actually have [2]. We thus seek answers to the following question:

RQ2: What were the effects of the quarantine, in terms of the quality of shared news articles, within `r/The_Donald`?

Finally, previous work evaluated interventions in isolation, even when these were part of a sequence. Similarly, many existing works evaluated intervention effects only within the moderated community, but an intervention might have, for instance, positive local effects but negative global ones (e.g., on other communities, or on the platform as a whole) [13, 28]. Studying intervention effects on other communities allows a thorough understanding of its consequences. Hence, we ask one final question:

RQ3: What were the effects of the sequence of interventions applied to r/The_Donald, on the other communities to which r/The_Donald members participated?

1.3 Summary of Methods, Findings, and Implications

1.3.1 Materials and Methods. Our observational study is based on Reddit data comprising more than 15M posts and spanning circa 2 years, collected from both r/The_Donald and other subreddits in which core users of the former participated.[2] We first operationalized intervention effects for user activity, comment toxicity, the degrees of both political polarization and factual reporting of shared news articles, and group community proclivity. We then leveraged appropriate statistical tests to assess all measured effects, including causal inference methods such as interrupted time series (ITS) regression analysis and Bayesian structural time series (BSTS) modeling.

1.3.2 Findings. Regarding the effects of the quarantine on activity and toxicity within r/The_Donald (**RQ1**), we find a moderate decrease in user activity as well as a strong short-term decrease, but a strong long-term increase in toxicity. The quarantine also caused a mild decrease in the degree of factual reporting of the news articles shared within r/The_Donald (**RQ2**). No significant effect was found for the political polarization of shared news articles. Results for **RQ1** and **RQ2** also reveal that core users of r/The_Donald suffered stronger effects with respect to other users. Finally, the analysis of the effects that the quarantine and the restriction had outside of r/The_Donald (**RQ3**) reveal a decrease in user activity, a marked increase in toxicity, a decrease on the degree of factual reporting, and an increase in political polarization of shared news articles. Results for **RQ3** also reveal that the restriction caused stronger effects than the quarantine.

1.3.3 Implications and Significance. Our results highlight that the sequence of interventions enforced on r/The_Donald had mixed effects. Overall, our results and other recent findings [28] partly question the positive judgments expressed in several previous works about the efficacy of Reddit's [10, 11, 45] and Twitter's [29] interventions. Furthermore, our nuanced results call for renewed efforts at evaluating the possible side effects of an intervention and the temporal variations of its effects. This research also contributes to building theories and methods [26, 31] to inform platform administrators for the design and deployment of future moderation interventions.

2 BACKGROUND AND RELATED WORK

We provide background information on r/The_Donald and the many issues emerged therein, and a critical discussion of the existing literature on the effects of past moderation interventions.

2.1 The Rise and Fall of r/The_Donald

Between 2015 and 2020 r/The_Donald served as an online space for supporters of the businessman and former US president Donald Trump. It was created on June 27, 2015, following the announcement of Trump's presidential campaign, and soon after it gained widespread popularity among Trump enthusiasts, as well as among conservative and libertarian users [34]. At its peak, it counted almost 800K subscribers and it frequently ranked in the top-10 subreddits by activity.[3] Due to their technical skills, organization and motivation [25, 30, 48], members of r/The_Donald managed to exert a strong influence on the news discussed on other social media platforms, like Twitter [10, 55].

Initially, discussions on r/The_Donald mainly focused on Trump-related news, with the vast majority of the posted content supporting his candidancy and presidency. However, through time the subreddit slowly regressed to an alt-right bastion [28, 49] and a hub for far-right extremism [10].

[2] The main data and code of the study are available at https://doi.org/10.5281/zenodo.6250576
[3] https://subredditstats.com/r/the_donald

The offensive nature of the content posted on r/The_Donald and the aggressive behavior of its members frequently caused considerable controversy and turmoil. Through the years, Reddit users, journalists and scholars repeatedly denounced r/The_Donald for being toxic and violent [28, 49], racist, sexist and Islamophobic [25, 44], engaged in coordinated trolling and harassment [22], in strategic manipulation [48], and in the spread of conspiracy theories [34]. The archetype of r/The_Donald's member was that of a white Christian male interested in conspiracy theories, firearms, and video games, and engaged in shocking and vitriolic humor [34].

Many of the aggressive and harmful behaviors described above were in clear violation of Reddit's policies. For this reason, between 2019 and 2020 the platform administrators applied three increasingly restrictive moderation interventions to r/The_Donald. The first of such interventions (i.e., the quarantine) applied concepts of design friction [14] in order to make it more difficult for casual users to enter, and to be exposed to the content of, the subreddit [10]. Apart from the added difficulties however, all of the content from r/The_Donald remained visible and all interactions with content and users remained possible. Conversely, the second intervention (i.e., the restriction) made it impossible for the vast majority of users to post new submissions to the subreddit and eventually resulted in a mass migration of users to a new platform [28]. In practice, the restriction doomed r/The_Donald, even before the final ban that occurred four months later. Our present work is part of the ongoing stream of research that aims at assessing the effects of these moderation interventions.

2.2 Evaluating the Effects of Moderation Interventions

The many issues that currently affect online platforms —including those described above, as well as the spread of mis- and disinformation [20, 50]; of propaganda and conspiracy theories [17, 39]; the rise of hateful, abusive, and coordinated inauthentic behavior [24, 38]; and the misbehavior of social bots and trolls [15, 35, 56]— mandate the design and deployment of a multitude of moderation interventions. Given this picture, a fundamental question arises about the efficacy of such interventions for mitigating the existing issues.

The recent body of works that evaluated Reddit's quarantines, restrictions, and bans represents the literature that is mostly related to our present study. Chandrasekharan et al. as well as Saleem and Ruths evaluated the effects of the bans that targeted r/fatpeoplehate and r/CoonTown in 2015, two subreddits whose users were known for harassment [11, 45]. Chandrasekharan et al. measured overall positive effects for the interventions. Specifically, they found that many former members of r/fatpeoplehate and r/CoonTown ceased using Reddit and that those who remained on the platform markedly decreased their hate speech usage. In addition, Saleem and Ruths found that the counteractions taken by the former r/fatpeoplehate members to circumvent the ban were short-lived and ineffective [45]. As it often happens with deplatforming, members who remained on Reddit after the bans "migrated" to other subreddits [28, 37]. Those subreddits, however, saw no significant changes in hate speech usage after the interventions [11]. Nonetheless, the former members of r/CoonTown more than doubled their posting activity when they migrated to r/The_Donald [10]. Reddit's quarantining of r/The_Donald and r/TheRedPill were evaluated in [10]. Authors concluded that the quarantines made it more difficult to recruit new members to the moderated communities, but that the overall degree of misogyny and racism of their existing members remained unaffected.

The previous studies shed light on (some of) the effects that Reddit's interventions had within Reddit itself. However, since such interventions caused many users to migrate to other platforms, those studies did not provide answers as to whether Reddit's bans made those problematic users «someone else's problem» [11]. Horta Ribeiro et al. aimed to answer this question in the aftermath of Reddit's 2020 deplatforming of r/The_Donald and r/Incels, whose former users migrated respectively to thedonald.win and incels.co [28]. They found that both interventions significantly decreased activity on the new platforms, reducing the number of shared posts, active users, and

Table 1. Overview of the 3 datasets used to answer research questions (RQ) regarding the interventions (I) quarantine (Q) and restriction (R) on r/The_Donald (TD) and its core users (CU).

Dataset	RQ	I	Submissions	Comments
TD	1 & 2	Q	981,980	11,400,674
CUw/iTD	1 & 2	Q	233,789	3,293,273
CUw/oTD	3	Q & R	148,054	3,191,170

newcomers. However, former users of r/The_Donald showed increases in toxicity and radicalization, supporting the hypothesis that the reduction in activity may have come at the expense of a more toxic and radical community [28]. Besides Reddit, Jhaver et al. evaluated Twitter's ban of three controversial influencers. They found that the deplatforming intervention significantly reduced the number of conversations about all three individuals and that their supporters decreased their overall activity and their degree of toxicity after the intervention [29].

2.2.1 Relation with Prior Work. Results from the studies previously discussed surface a general consensus that deplatforming interventions lead to overall positive results, at least with respect to reducing activity. At the same time however, nearly all studies also reported a number of negative side effects. In light of these results, one of the most recent studies along this line of research underlined the importance of carrying out nuanced analyses in order to accurately assess the (sometimes mixed) effects of moderation interventions applied to complex online social systems [28]. Within this scientific context, our present work extends and complements existing studies by investigating a number of unexplored aspects. Firstly, we extend previous analyses on activity and toxicity by also investigating the quality of the news articles (i.e., in terms of their political polarization and the degree of factual reporting) shared and consumed —important aspects that contribute to accurately understanding the consequences of moderation interventions, including their negative side effects [2]. Secondly, we compare the effects of subsequent interventions applied to r/The_Donald, thus contributing to analyze their relative effectiveness [10, 28]. Finally, we contribute to filling the scientific gap related to the analysis of Reddit's quarantines, whose effects were so far only analyzed by [10] within the moderated community. Here instead, we also evaluate the nuanced effects that quarantining r/The_Donald had on other communities on Reddit, similarly to what [28] did when moderated communities migrated to other platforms.

3 DATA

Our study is based on observational Reddit data comprising more than 15M posts collected over the course of two years. In detail, our data is organized into three datasets that respectively contain: (i) all content shared within r/The_Donald in a time frame centered around the quarantine; (ii) all content shared within r/The_Donald by core users of the subreddit, in a time frame centered around the quarantine; and (iii) all content shared outside of r/The_Donald (i.e., in all other subreddits) by core users of r/The_Donald, in a large time frame that encompasses both quarantine and restriction. The mapping of our datasets to our research questions is shown in Table 1, together with other summary information. In detail, we use the first and second dataset to answer to **RQ1** and **RQ2** (quarantine effects within r/The_Donald). Instead, we use the third dataset to answer to **RQ3** (quarantine and restriction effects outside of r/The_Donald).

Posts on Reddit can be either submissions or comments. These represent two intrinsically different activities that, in general, follow different dynamics [53]. Moreover, certain moderation interventions (e.g., the restriction suffered by r/The_Donald) are designed to have a strong impact on

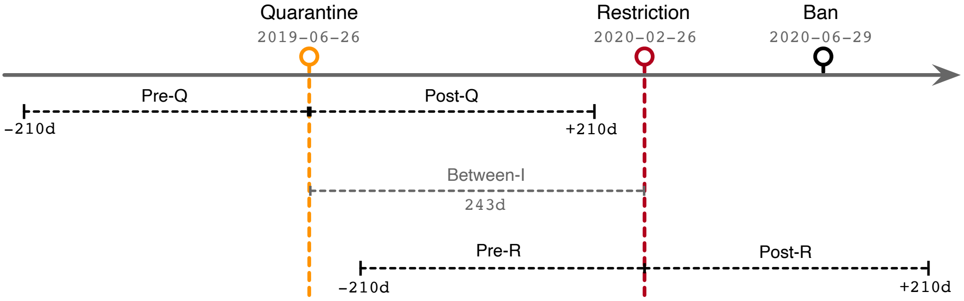


Fig. 2. Timeline depicting the pre-post periods used for data collection and analysis, based on 2 time windows centered around the quarantine and the restriction, plus a convenience period between both interventions.

submissions, while leaving commenting activities unaffected. We thus conducted separate analyses for submissions and comments.

3.1 Data Source and Time Frame

For the collection of all Reddit data used herein, including the posts from r/The_Donald that are no longer available in Reddit because of the ban, we used the monthly archives from Pushshift, a service that provides historical Reddit data [3]. As already shown in Figure 1 in §1.1, daily posting activity in r/The_Donald came to a halt shortly after the restriction, thus the subreddit was already inactive when the ban occurred. Consequently, and in line with previous work [28], the time frame considered for our data collection and analyses is centered around the quarantine and the restriction. In literature, different time frames were used: ±60 days around interventions [45], ±120 days [28], ±180 days [10, 29], and ±200 days [11]. Given that circa 245 days passed between the quarantine and the restriction, we used a time frame spanning ±210 days (30 weeks). This choice allows us to divide a given time window around an intervention by groups of ten days [10, 11] or seven days, which eases the analysis of our time series. We thus collected data from November 28, 2018 to September 23, 2020 (inclusive), in which 1.06M submissions and 12.3M comments were posted to r/The_Donald. We then further divided the daily content into two pre-post intervention periods around quarantine (Q) and restriction (R), plus a convenience period between both —all of which exclude the intervention days— as illustrated in Figure 2.

3.2 Core Users

The selection of representative members of a community (i.e., core users) is an inherently subjective —but sometimes inevitable [38]— process. Before the quarantine r/The_Donald was a public space in which any registered user could post. As a result, there are many users who participated in the subreddit very sporadically, or even only once during the time frame of our study. Moreover, several users authored many posts in r/The_Donald, but only on a single day, or for a single submission or thread. Keeping in mind that interventions can have different effects depending on user characteristics and that the interventions enforced by Reddit were targeted at the core members of r/The_Donald, we operationalized our definition of *core users*. We identify core users and we define them as those users who authored at least one post (i.e., either a submission or comment) a week, for the whole 30 weeks of the pre-quarantine period. In this way, we ensure that each core user had a non-negligible posting activity and a prolonged involvement with the community (circa seven months), before any intervention took place. Based on this definition, and after removing a few moderation bot accounts, we identified 2,239 core users of r/The_Donald,

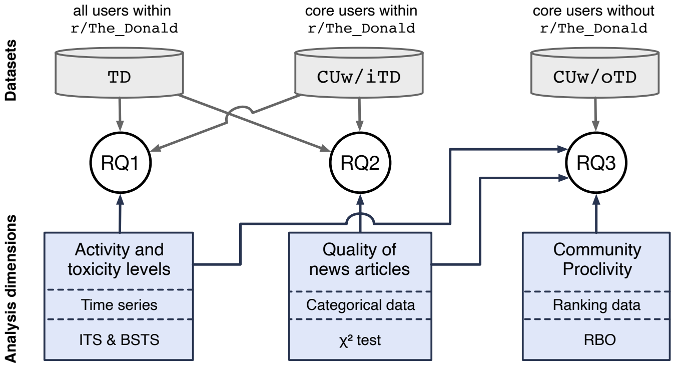


Fig. 3. Conceptual approach used herein to investigate r/The_Donald interventions' community effects on activity and toxicity (RQ1), on the quality of its shared news articles (RQ2), and on other subreddits (RQ3).

which represent only 1.12% of the 198.5K distinct post authors in the subreddit for the study time frame. Despite their small number, core users posted 37.5% of the total submissions and more than 53.9% of the total comments.

3.3 Datasets Construction and Preliminary Analysis

After the selection of the appropriate study time frame (§3.1) and of the core users (§3.2), we constructed our three datasets, summarized in Table 1. The first dataset (TD) is centered around the quarantine, and contains r/The_Donald posts made by all users (982K submissions and 11.4M comments). The second dataset (CUw/iTD) is a subset of the previous, with only posts made by core users within r/The_Donald (233.8K submissions and 3.29M comments). The third dataset (CUw/oTD) contains all posts that core users made outside of r/The_Donald, independently on the subreddit. This last dataset encompasses both the quarantine and the restriction.

In order to better inform our choice of inference methods we also conducted an exploratory time series analysis on daily content and unique users for all of the datasets. This step allowed us to test for autocorrelation and seasonality before any intervention (i.e., Pre-Q period), which can cause certain causal inference methods to yield inaccurate results [7]. For autocorrelation, we applied the Durbin-Watson (DW) test of the lmtest R package. To identify seasonality, we used iterative STL (seasonal and trend decomposition using LOESS) of the forecast R package, with which we also derive seasonal strength on a scale from zero to one. All of the time series analyzed showed a significant positive autocorrelation, with values for the DW statistic ranging from 0.582 to 1.536. We also identified a weekly seasonality, with a strength that ranges from modest to strong (0.181–0.764). Both general and core users of r/The_Donald were most active around mid-week and least active around the weekend during the Pre-Q period, albeit core users showed higher weekly seasonal strength. This seasonality also indicates that having a subdivision of time periods in groups of seven days is more appropriate than the ten days used in other works [10, 11].

4 METHODS

Based on the nature of our datasets, we defined intervention effects, gathered additional data, and selected statistical descriptive and inference techniques apt to answer to our research questions. The conceptual approach we followed is illustrated in Figure 3 and detailed in the following paragraphs.

4.1 Operationalizing Effects

A single intervention can have multiple effects [33]. Therefore, in order to accurately assess intervention effects it is necessary to investigate possible changes in many dimensions.

4.1.1 Activity. We computed two separate content activity metrics based on Reddit's submissions and comments: (i) daily number of submissions and (ii) daily number of comments. Another commonly used activity metric is the number of daily active users (DAU). However, what constitutes an active user differs from platform to platform. We reprise the definition for core users given in §3.2 to define DAU as those users who performed at least a posting activity on a given day. Analogous to our metrics for posted content, we have (iii) submission DAU and (iv) comment DAU. These are thus the four metrics we use herein to measure intervention effects on activity.

4.1.2 Toxicity. Many previous studies evaluated the effectiveness of interventions for hate speech reduction [10, 11, 45]. However, we lack a common definition of hate speech [18, 44], its detection is often based on the presence of certain hateful words, with dictionary construction being challenging and context-dependent [18, 23], and the mere presence of hateful words has limited power due to the varied expressions of online incivility. For these reasons, others recently used toxicity as a more general concept to study moderation interventions [28, 29] and as a discussion quality indicator [41]. In this context, the Jigsaw unit at Google developed the Perspective API,[4] a widely used public service that computes multiple toxicity-related scores for comments [43].

Herein we rely on the *severe toxicity* score of this service, similarly to several previous works [28, 29]. Severe toxicity is defined as being «very hateful, aggressive, disrespectful [...] or otherwise very likely to make a user leave a discussion or give up on sharing their perspective», and is considered to be the most reliable and robust[5] indicator of toxicity among those provided by the Perspective API [28, 41]. From severe toxicity, defined in the $[0, 1]$ range, we compute two metrics of toxicity for Reddit comments: daily median and daily relative frequency. The first one is computed as the median of the severe toxicity scores of comments aggregated on a daily basis. For the second metric, we first convert scores into binary labels by considering as "severely toxic" all those comments whose score is above a certain threshold (≥ 0.5), we then compute the fraction of severely toxic comments on a daily basis. We include both metrics in our analyses because they provide slightly different information and both have been recently used in studies on online toxicity [28, 29, 41]. For the second metric we chose a score ≥ 0.5 because we consider it a good sensitive default for binary classification, it has been used in previous literature [41], and results are qualitatively similar with higher scores used in other related work, such as ≥ 0.8 [28].

4.1.3 Political Polarization. The degree of polarization of a community is another important metric when investigating the "health" of online spaces, with several interventions being specifically designed to reduced online polarization [2]. Still, concerns have surfaced about the possibility that deplatforming interventions might actually increase the polarization of affected users [28]. For this reason we also evaluate intervention effects in terms of the overall degree of political polarization of the affected users. To measure political polarization, we adopt an approach based on the analysis of news articles shared by users. To this end, we leverage data from Media Bias/Fact Check (MBFC),[6] a widely-used platform that provides expert-curated fact checks and audit information about a large set of US media outlets [6]. Despite the limitations of using a single source for bias rating, the validity of using MBFC data to study political bias and factual reporting of Reddit submissions has been demonstrated via a large-scale study with hundreds of millions of posts across thousands

[4] https://www.perspectiveapi.com/
[5] The severe toxicity score in Perspective is specifically designed to avoid predicting benign usage of foul language as toxic.
[6] https://mediabiasfactcheck.com/

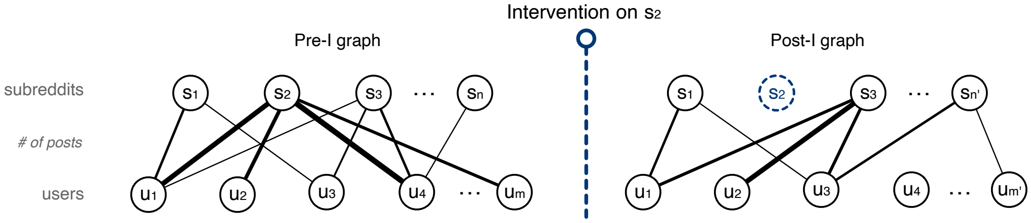


Fig. 4. Representation of the user-subreddit relationship as a weighted undirected bipartite graph for the pre-post periods around an intervention on a given subreddit. As shown, an intervention on s_2 might induce users to alter their overall participation habits, thus altering the group community proclivity.

of subreddits [52]. Therefore, we first use the general *bias* in MBFC data, in which news websites are classified as *pro-science* (factual and non-politically inclined), *satire*, *conspiracy/pseudoscience* (unverifiable information related to known conspiracies), *fake* (deliberate attempts to publish hoaxes and/or disinformation for profit or influence), or *politically biased*. For the latter, MBFC provides a Likert-like rating with five labels based on the US political spectrum, ranging from “left” to “right”, which we then use to measure political polarization. In our analyses we compare the distribution of such labels for the articles shared by a community before and after a given intervention, treating general bias as categorical data and political bias as ordinal data.

4.1.4 Factual Reporting. Grounding comments on facts and sharing a common vision of reality with those with which we interact are key elements to achieve healthy online conversations [4]. Because of this, scholars have long monitored the extent to which online communities share and consume news from sources known for adopting fact-based journalistic practices [6]. Herein, we evaluate the extent to which users affected by a moderation intervention change their behavior with respect to the degree of factuality of the news sources they share, by leveraging once again data from MBFC. Specifically, for each news outlet MBFC provides a *factual reporting* Likert-like rating with six labels, ranging from “very low” to “very high” factuality. Similarly to our analyses of political polarization, we use factual reporting to study intervention effects by comparing their distribution before and after a given intervention, treating the labels as ordinal data.

4.1.5 Group Community Proclivity. User participation in online communities can be represented not just by the number of posts or unique users (what we herein call activity), but also by the overall structural relationship between a group of users and the set of communities in a platform. That is, when a group of users —as a whole— is more inclined to participate in certain communities rather than others, the former are more important for that group. We refer to this concept as *group community proclivity*. We are thus interested in comparing group community proclivity of core users of `r/The_Donald` within Reddit before and after interventions on the subreddit. Since we carry out separate analyses for submissions and comments, we have two bipartite graphs for each analyzed period. Our approach allows us to compare structural changes in the user-subreddit relationship before and after a given intervention, as illustrated in Figure 4.

To obtain a metric of group community proclivity, we first represent the user-subreddit relationship for a given period as a weighted undirected bipartite graph $G = (U, S, E, w)$, where U is the set of core users, S is the set of subreddits, E is the set of edges connecting nodes in U and S, and w are the weights of edges. An edge $e_{ij} \in E$ exists from $u_i \in U$ to $s_j \in S$ if u_i has authored content on s_j in the given period, with $w_{ij} \in \mathbb{N}^+$ being the number of posts authored by u_i in s_j. We then adopt a ranking algorithm for bipartite graphs so as to obtain a ranking of subreddits according to their importance for the group of users. Among the available ranking algorithms, we adopted Co-HITS —a version of the well-known HITS algorithm adapted to bipartite graphs [19]. In our analyses we used the Co-HITS implementation from the `birankr` R package.

4.2 Descriptive and Inference Methods

We now describe our choice of statistical methods for assessing intervention effects.

4.2.1 RQ1: Effects on Activity and Toxicity within r/The_Donald. We followed a quasi-experimental approach to measure the intervention effects on activity and toxicity within `r/The_Donald`. Based on our preliminary time series analysis (§3.3), we leverage two causal inference methods to analyze possible intervention effects in a complementary manner: interrupted time series (ITS) regression analysis to describe the trend, onset and decay of intervention effects; and Bayesian structural time series (BSTS) modeling to compute effect size, confidence intervals, and significance.

ITS regression analysis aims to establish the underlying trend of the variable of interest across a continuous sequence of observations before and after being "interrupted" by an intervention at a well-defined point in time. In its most basic form, ITS can be implemented through simple segmented linear regression models, which renders it easy to use, understand, and plot. For these reasons it was used in works on social media interventions, either as complementary to simpler methods such as difference-in-differences [11], or as the only causal inference method [10, 29]. However, specific regression families should be used in the case of non-linear trends and for particular underlying data distributions [5]. Moreover, ITS has caveats in case of autocorrelation and seasonality, which demand adequate model adjustments so as to avoid misleading estimates of effect sizes, using for instance apt autoregressive integrated moving average (ARIMA) models [46], or heteroscedasticity and autocorrelation consistent (HAC) estimators [27, §15.4]. Herein we use a simple and interpretable ITS regression to visualize variations in the linear trends of the variables of interest around interventions, according to the following segmented linear model,

$$Y_t = \beta_0 + \beta_1 T + \beta_2 D + \beta_3 P + \epsilon \tag{1}$$

where Y_t is the outcome variable of the time series; T is a continuous variable that indicates the time in days from the start of the observational period, with β_1 indicating the trend before the intervention; D is a dummy variable indicating the presence (1) or absence (0) of the intervention, with β_2 indicating the immediate change upon intervention; P is a continuous variable that indicates the days passed since the intervention has occurred (from 0 to 210), with β_3 indicating the change in trend after intervention; finally, ϵ is the error term of the model.

BSTS, on the other hand, is an approach based on Bayesian statistics that uses a structural time series model to capture the trend, seasonal, and related components of a time series, together with a dynamic regression component using Monte Carlo Markov Chain (MCMC) to create counterfactual data and confidence intervals [47]. BSTS improves on ITS in two main aspects: it provides a fully Bayesian estimate for the effect across time, which can be updated as new information is available; and it uses model averaging to construct a synthetic control to model the counterfactual. This means that the model's certainty will increase in case additional information is available, and that we can create a useful model even in the absence of an external control group. This is particularly valuable for our analyses due to the difficulties in finding an independent subreddit to act as a control. Indeed, Reddit users are mostly free to participate in as many subreddits at the same time as they wish, with most doing so, especially across related communities. This results in a significant user overlap in many communities, which thus are not independent from each other. Additionally, before the quarantine `r/The_Donald` was one of the most active subreddits, despite its focus on a single political figure. Finding a "control" subreddit with similar characteristics, let alone an independent one, would be problematic [10]. For these reasons, and despite the adoption of ITS in many recent studies that measured intervention effects [10, 11, 29], BSTS represents a better alternative for estimating effects, given also the autocorrelation and weekly seasonality of our data. For our analyses we use the BSTS implementation of the `CausalImpact` R package.

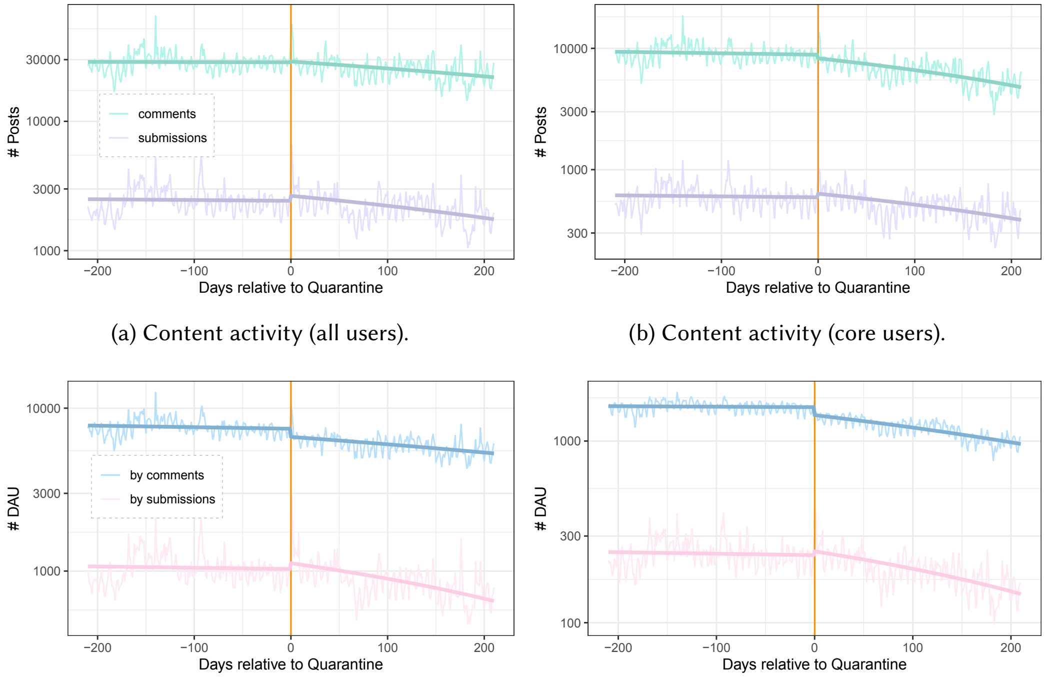
4.2.2 RQ2: Effects on the Quality of Shared News Articles within r/The_Donald. Concerning the possible changes in the quality of shared news articles in `r/The_Donald`, we compare the political polarization and factual reporting scores of the news outlets linked in Reddit submissions, before and after a given intervention. Considered that only circa half of the collected submissions contain a link to an external website, of which only a subset point to news outlets in the MBFC database, we aggregate these links by pre- and post-intervention periods. To probe effects in political polarization, we further aggregate data on the left and right side of the political spectrum (excluding neutral news outlets, which are assigned the "least biased" label by MBFC). Then, we compare the differences in the distributions of political polarization scores before and after an intervention. Finally, we apply χ^2 tests for significance, albeit with the awareness that the partially-matched data between interventions could give unreliable results in case of small differences. For estimating effects in factual reporting, we follow a similar approach where we first aggregate data by lower and upper ends on the factuality spectrum, and then we conduct χ^2 tests on the proportions.

4.2.3 RQ3: Effects of Multiple Interventions outside of r/The_Donald. To answer **RQ3**, we follow the methods to measure the possible effects on user activity and toxicity described for **RQ1**, as well as the methods for quality of shared news articles described for **RQ2**. In addition, we also evaluate the effects in group community proclivity (i.e., ranking variations computed with Co-HITS). In line with the approach used in RQ2, to assess group community proclivity we aggregate data on three periods: pre-quarantine, between-interventions, and post-restriction. Between-interventions is a convenience period that covers post-quarantine and pre-restriction, which we confirmed bears no significant difference with respect to the periods it substitutes, as there is less data compared to *activity* and their spans mostly overlap, as sketched in Figure 2. To measure proclivity change between two periods, we use the rank-biased overlap (RBO) for indefinite lists [51]. RBO is a probabilistic model of similarity between two ranking lists based on average overlap, with a bias in the proportional overlap at each depth of weights, which can handle tied ranks and rankings of different lengths [51]. The latter is important in our case as other methods used for ranking similarity (e.g., Kendall's tau distance) only handle lists with the same items. However, with time some subreddits become inactive and new ones are created. In addition, we are particularly interested in changes about the top subreddits with the most proclivity, not on whole lists. We compute RBO scores with the corresponding function of the `gespeR` R package.

5 RESULTS

5.1 RQ1: Effects on Activity and Toxicity within r/The_Donald

5.1.1 Activity. The ITS analysis (Eq. 1 in §4.2.1) for pre-quarantine shows that the linear trend (β_1) of activity in `r/The_Donald` was stable —albeit slightly decreasing— in terms of daily submissions, comments, and active users, as shown in Figure 5. Upon quarantine, there were activity spikes in all metrics that lasted circa two days, but in all cases there is a noticeable subsequent effect (β_3) of decline in activity. However, the ITS immediate effect (β_2) was heterogeneous in direction among the metrics. For instance, upon quarantine there was an immediate increase on submission DAU but a decrease on comment DAU, as visible in Figures 5c and 5d. At first glance, ITS seems to indicate that core users had a more marked decline in activity compared to all of the users. The BSTS analysis, reported in Table 2, shows that all activity metrics had significant negative relative



(a) Content activity (all users).

(b) Content activity (core users).

(c) Daily active users (all users).

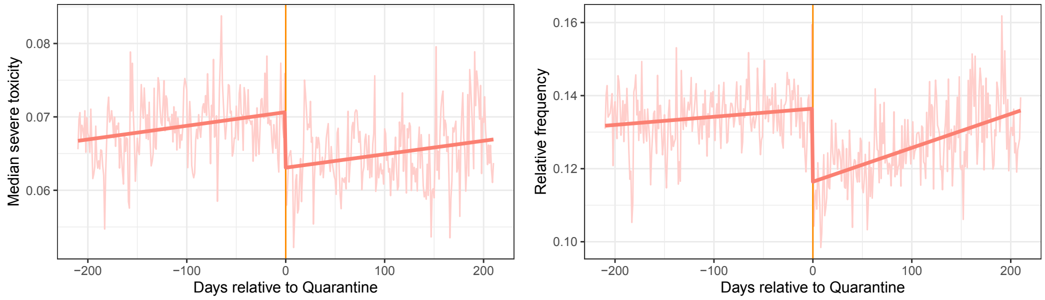
(d) Daily active users (core users).

Fig. 5. ITS regressions of activity within r/The_Donald around quarantine, for all users and core users.

Table 2. BSTS results of quarantine effects on activity within r/The_Donald for all users (TD) and core users (CUw/iTD). The lower the posterior tail-area probability (*pp*), the higher the probability of a causal effect.

	Post-intervention avg. value			Relative effect (%)		
	Actual	Predicted	95% CI	% Change	95% CI	*pp*
TD submissions	2201	2406	[2215, 2609]	−8.5	[-17, -0.57]	.019
TD comments	25457	28501	[27019, 29879]	−10.7	[-16, -5.5]	.001
TD submission DAU	886	1021	[941, 1091]	−13.2	[-20, -5.4]	.001
TD comment DAU	5967	7492	[7176, 7781]	−20.4	[-24, -16]	.001
CUw/iTD submissions	509	589	[545, 631]	−13.6	[-21, -6.2]	.001
CUw/iTD comments	6529	8940	[8418, 9400]	−27	[-32, -21]	.001
CUw/iTD submission DAU	196	235	[222, 248]	−16.7	[-22, -11]	.001
CUw/iTD comment DAU	1176	1529	[1493, 1562]	−23	[-25, -21]	.001

effects (i.e., average difference between observed and predicted post-intervention values). The most important effects concerned the number of daily active core users (-23%) and the respective comments (-27%), whilst the least important were on posts made by all users (-8.5% for submissions and -10.7% for comments). Hence, the quarantine indeed had a highly significant effect of reducing the activity of core users within r/The_Donald.



(a) Median severe toxicity scores.

(b) Fraction of severely toxic comments.

Fig. 6. ITS regressions of core users toxicity within r/The_Donald around the quarantine.

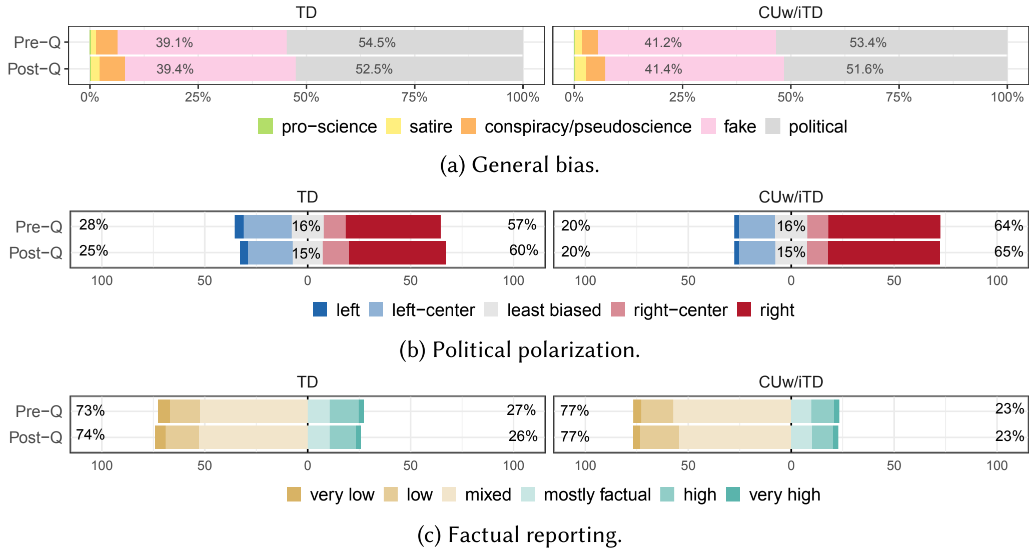
Table 3. BSTS results of quarantine effects on core users toxicity within r/The_Donald. The lower the posterior tail-area probability (*pp*), the higher the probability of a causal effect.

	Post-intervention avg. value			Relative effect (%)		
	Actual	Predicted	95% CI	% Change	95% CI	*pp*
Median severe toxicity	.065	.069	[.068, .070]	−6	[-7.9, -4]	.001
Severely toxic comments	.126	.136	[.132, .137]	−6.4	[-8.2, -4.8]	.001

5.1.2 Toxicity. We collected toxicity scores for 3.29M comments made by core users within r/The_Donald around the quarantine, which represent 32.1% of the non-erased comment total for the same time span. Based on the ITS analyses shown in Figure 6, during pre-quarantine the median severe toxicity score of core users' comments had a noticeable increasing linear trend. However, upon quarantine there was a significant immediate downward effect, with the median reaching its lowest value a few days later. Despite this drop in toxicity, the previous increasing trend reprised during post-quarantine, reaching median levels similar to those in pre-quarantine. Regarding "severely toxic" comments, there is a less marked increasing linear trend compared to the median, although both the immediate effect and the subsequent increase post-quarantine are much more evident. Additional analyses on the robustness of our results confirmed that these effects hold at different thresholds of severe toxicity used for labeling "severely toxic" comments, as well as with other Perspective toxicity scores (e.g., toxicity, insult, threat). According to the BSTS results reported in Table 3, the causal effects can be considered statistically significant (posterior tail-area probability $pp \approx .001$), but as previously stated, there is a noticeable decay of the effects, with the relative frequency of comments classified as "severely toxic" reaching even higher levels compared to pre-quarantine. The quarantine thus had the strong immediate effect of reducing toxicity, but also the strong long-term effect of increasing it.

5.2 RQ2: Effects on the Quality of Shared News Articles within r/The_Donald

5.2.1 Political bias. Most news articles (53%) shared within r/The_Donald came from politically biased sources, as shown in Figure 7a, with the majority of these (62%) falling at the right of the US political spectrum, as shown in Figure 7b. News shared by core users were more biased to the right, compared to all users, with 64% and 58% respectively. After quarantine, there was a small (3%) but significant increase in polarization to the right for all users ($\chi^2 = 90.3; p < .001$), while for core users there was a non-significant increase of only 1% ($\chi^2 = 0.641; p = .423$).

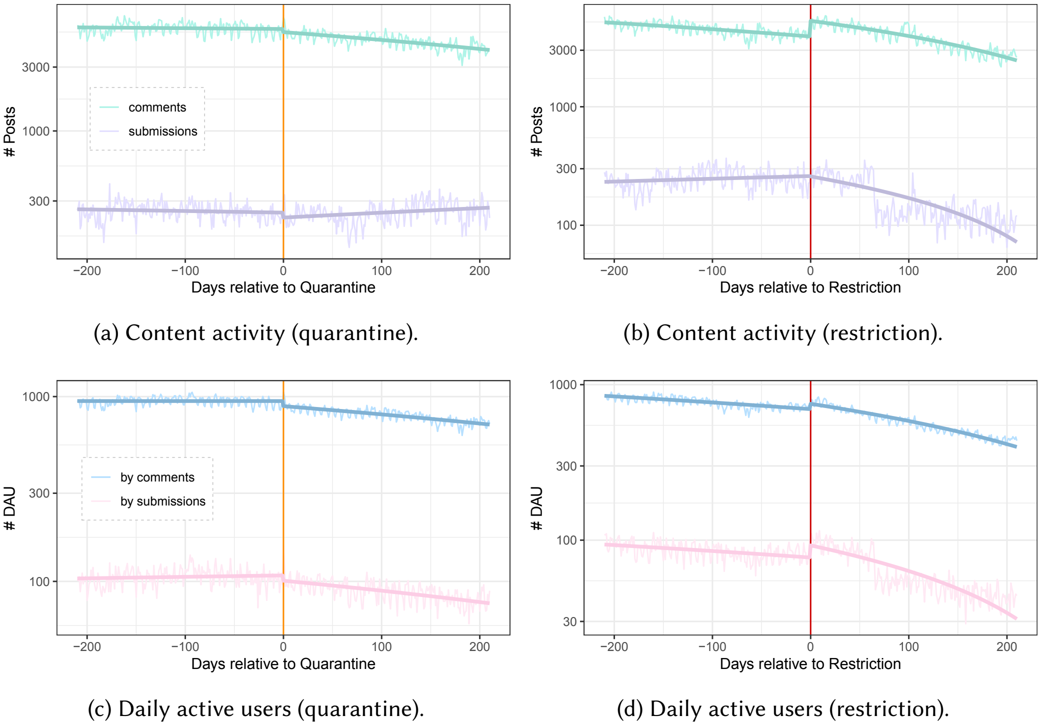


(a) General bias.

(b) Political polarization.

(c) Factual reporting.

Fig. 7. Quality of news outlets in r/The_Donald around quarantine for all users (TD) and core users (CUw/iTD).



(a) Content activity (quarantine).

(b) Content activity (restriction).

(c) Daily active users (quarantine).

(d) Daily active users (restriction).

Fig. 8. ITS regressions of core users activity outside of r/The_Donald around quarantine and restriction.

Table 4. BSTS results of quarantine (Q) and restriction (R) effects on core users activity outside r/The_Donald. The lower the posterior tail-area probability (*pp*), the higher the probability of a causal effect.

	Post-intervention avg. value			Relative effect (%)		
	Actual	Predicted	95% CI	% Change	95% CI	*pp*
Q submissions	246	249	[237, 260]	−0.9	[-5.4, 3.7]	.343
Q comments	4748	5780	[5610, 5948]	−17.9	[-21, -15]	.001
Q submission DAU	88	106	[102, 109]	−16.2	[-20, -13]	.001
Q comment DAU	797	932	[909, 950]	−14.5	[-16, -12]	.001
R submissions	165	247	[234, 263]	−33.2	[-39, -28]	.001
R comments	3886	4161	[3835, 4472]	−6.6	[-14, 1.2]	.051
R submission DAU	62	82	[78, 86]	−24.6	[-30, -19]	.001
R comment DAU	575	725	[7176, 7781]	−20.6	[-26, -15]	.001

5.2.2 Factual reporting. Figure 7c shows that around 75% of shared news content came from outlets in the lower half of the factual reporting spectrum, with 40% of the total shared articles deemed as mostly *fake* (see definition in §4.1.3). On average, core users shared slightly less reliable content compared to all users, but there was no significant change in factual reporting between pre- and post-quarantine ($\chi^2 = 1.0353; p = .308$), whereas the factuality decrease for all users was small (-1.2%) but significant ($\chi^2 = 47.37; p < .001$).

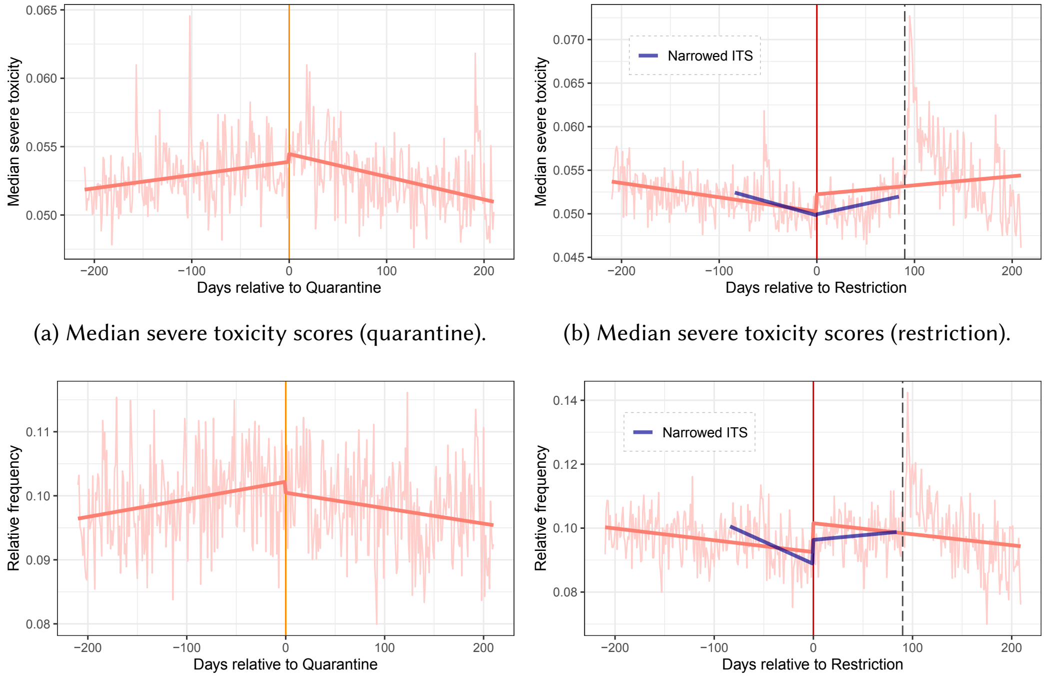
5.3 RQ3: Effects of Multiple Interventions outside of r/The_Donald

5.3.1 Activity. In general, core users activity outside of r/The_Donald suffered an immediate and sustained decrease in the post-quarantine period (consistent with activity within the subreddit), except for the number of daily submissions, which manifested an increasing trend post-intervention. This means that the remaining active core users increased their activity in terms of submissions to other subreddits, after a slight decrease upon quarantine. Indeed, the BSTS analysis reported in Table 4 indicates that there is no significant effect for submissions after quarantine (relative effect of -0.9%, $pp = .343$), whilst for the other metrics we measured significant effects ($pp = .001$), with relative effects ranging from -14.5% to -17.9%.

The situation changes in the post-restriction period. On the one hand, even if the number of DAU that commented significantly decreased (relative effect of -20.6%, $pp = .001$), the number of comments had an immediate slight increase upon restriction and a less pronounced decreasing trend (relative effect of -6.6%, $pp = 0.051$), as illustrated in Figures 8b and 8d. On the other hand, the restriction had a strong and significant effect ($pp = .001$) on both the number of submissions (-33.2%) and submission DAU (-24.6%). In part, this decrease is likely due to the migration of members of r/The_Donald to a different platform (thedonald.win) upon restriction.

5.3.2 Toxicity. We collected toxicity scores for 3.19M comments made by core users outside of r/The_Donald, which is a lower number with respect to the 3.5M comments made within r/The_Donald by the same users during the same time frame. Regarding the quarantine, the pre-intervention periods of both median values and relative frequency of severely toxic comments had a moderate increasing linear trend, which upon quarantine became decreasing, as shown in Figures 9a and 9c. The BSTS results summarized in Table 5 show that the effect for the median values was not significant ($pp = .125$), while it is significant for severely toxic comments ($pp = .021$).

Concerning the restriction, the situation is more complex, as there is a remarkable surge in toxicity during the weeks of the George Floyd protests, started on May 26, 2020. These protests had a marked



(a) Median severe toxicity scores (quarantine).

(b) Median severe toxicity scores (restriction).

(c) Fraction of severely toxic comments (quarantine).

(d) Fraction of severely toxic comments (restriction).

Fig. 9. ITS regressions of core users toxicity outside of r/The_Donald around the quarantine and the restriction. A vertical dashed line indicates the start of the George Floyd protests, after which toxicity surged. A narrowed (±12 weeks) ITS regression analysis (blue-colored) excludes the protests.

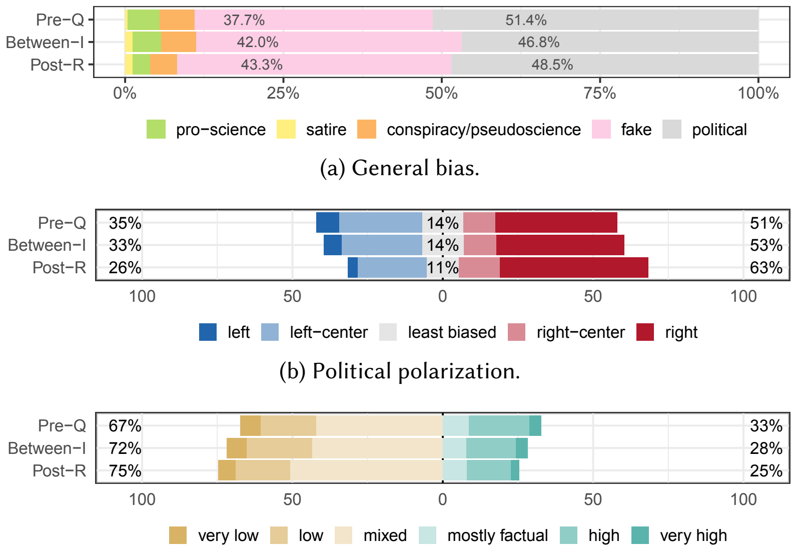
polarizing effect in the US public opinion, particularly between liberals and conservatives [42]. Our result is consistent with the difficulties encountered by Horta Ribeiro et al. in analyzing toxicity in r/The_Donald. For these reasons, we performed additional analyses with narrowed pre-post restriction periods of ±12 weeks, thus excluding these protests. In the pre-restriction span, the median values and relative frequencies showed a decreasing linear trend, clearly visible in Figures 9b and 9d. All of the post-restriction trends show an upward trend, except for the 30-week period pertaining to relative frequency of toxic comments in which is downward. Based on the BSTS analysis, whose results are in Table 5, post-intervention effects are significant ($pp \leq 0.003$), except for the median values during quarantine ($pp = .125$) and narrowed restriction ($pp = .462$).

5.3.3 Political bias. During the time frame of the study, there were 399K submissions made by core users outside of r/The_Donald, of which 91K have an external link (22.9%), and 49K are present in MBFC (12.37% of the submission total). Around half of the shared news links pertained to politically biased outlets. Before any intervention, core users shared fewer articles from outlets biased to the right of the political spectrum outside of r/The_Donald (51%) compared to within (64%). However, upon both quarantine and restriction there was a significant increase in bias to the right in content posted in other subreddits, respectively reaching 53% ($\chi^2 = 13.07; p < .001$) and 63% ($\chi^2 = 95.8; p < .001$) in the post-interventions periods, as shown in Figure 10b.

5.3.4 Factual reporting. We noticed a similar phenomenon in the decrease of factuality of shared news after each intervention, as it can be seen in Figure 10. Moreover, before any intervention,

Table 5. BSTS results of quarantine (Q) and restriction (R) effects on core users toxicity outside `r/The_Donald`. Pre-post periods span 30 weeks, except those narrowed (n.) to 12 weeks to exclude the George Floyd's protests. The lower the posterior tail-area probability (*pp*), the higher the probability of a causal effect.

	Post-intervention. avg. value			Relative effect (%)		
	Actual	Pred.	95% CI	% Change	95% CI	*pp*
Q Median severe toxicity	.052	.053	[.052, .054]	−0.73	[-2, 0.45]	.125
Q Severely toxic comments	.098	.100	[.098, .102]	−2.1	[-4.3, -0.13]	.021
R Median severe toxicity	.053	.051	[.050, .052]	+3.9	[2.5, 5.6]	.001
R Severely toxic comments	.098	.095	[.093, .097]	+3.3	[1.3, 5.7]	.003
R (n.) Median severe toxicity	.053	.051	[.050, .052]	−0.14	[-1.7, 1.4]	.462
R (n.) Severely toxic comments	.098	.094	[.092, .097]	+3.5	[0.8, 6.1]	.004



(a) General bias.

(b) Political polarization.

(c) Factual reporting.

Fig. 10. Quarantine (Q) and restriction (R) effects on the quality of news articles shared outside of `r/The_Donald` by core users.

core users shared more factual content outside of `r/The_Donald` compared to within, as visible in Figures 7c and 10c. For instance, in the pre-quarantine period the share of news articles from unreliable sources was 67% outside of `r/The_Donald` and 73% within. Between interventions, the low-factuality outside of `r/The_Donald` increased to 72% ($\chi^2 = 90.1; p < .001$), and in the post-restriction period increased to 75% ($\chi^2 = 28.8; p < .001$). This change is most likely related to the increase in proportion of articles from *fake* outlets after each restriction, as shown in Figure 10a, which during pre-quarantine period was 37.7%, then 42% between the restrictions, and finally 43.3% in the post-restriction period.

5.3.5 Group community proclivity. Results of intervention effects on group community proclivity denote a different proclivity dynamic for submissions and comments, as reported in Table 6. Specifically, the rankings of most prominent subreddits based on submissions are less similar

Table 6. Rank-biased overlap (RBO) between group community proclivity (top 50 subreddits) by content kind and paired intervention periods. The higher the RBO scores, the more similar the proclivity between periods.

	Proclivity RBO	
	Submissions	Comments
Pre-Q & Between-I	.329	.658
Between-I & Post-R	.299	.431
Pre-Q & Post-R	.145	.354

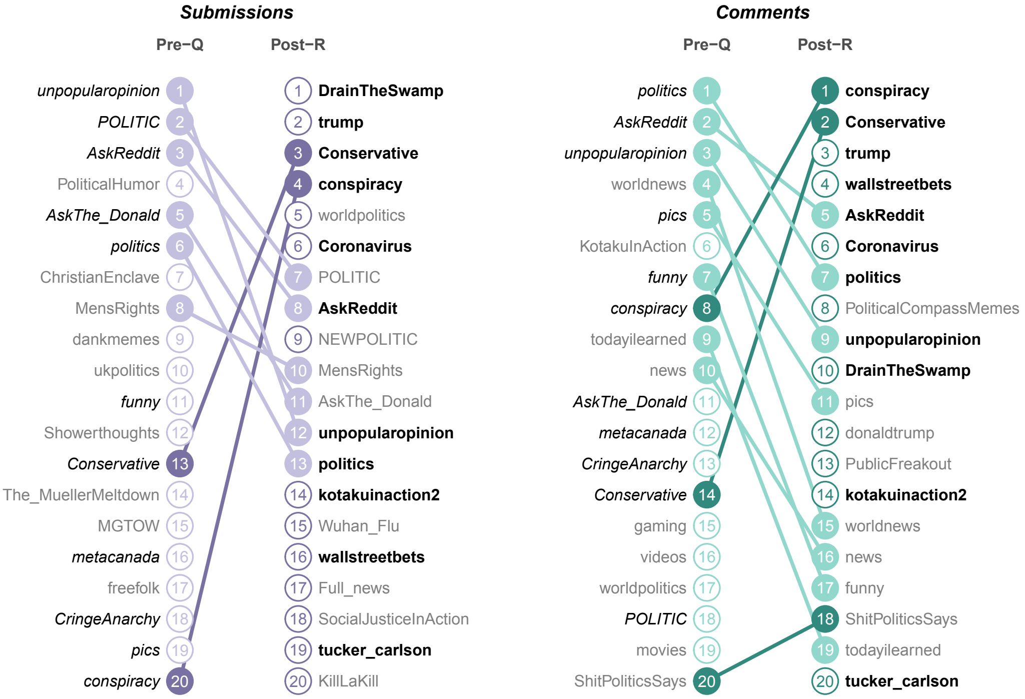
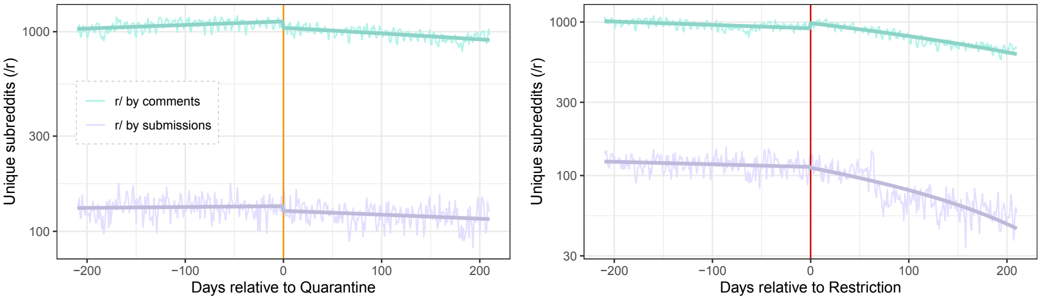


Fig. 11. Top 20 subreddits ranked by group community proclivity. Subreddits in italic and bold are present in both content lists in Pre-Q and Post-R, respectively. Dark-colored subreddits increased their ranking after the intervention, while lightly-colored subreddits decreased their ranking after the intervention.

among intervention periods compared to those based on comments. In particular, the change in similarity for Pre-Q and Between-I by submissions (.329) is much lower than the respective value by comments (.658). Thus, after the quarantine, core users showed a higher proclivity for a much different set of subreddits in which to post submissions, compared to comments. This phenomenon can also be seen in the changes in proclivity between Pre-Q and Post-R periods depicted in Figure 11, in which the subreddits that are present in the top 20 lists for both periods, are fewer for the lists based on submissions than for those based on comments. Incidentally, three subreddits climbed to the top-4 positions in both content types during Post-R: `r/conspiracy`, `r/Conservative`, `r/trump`.

It should be noted that these changes in proclivity could have been influenced by changes in the number of subreddits in which core users participated. Hence, we also investigated the effect of interventions on the number of daily distinct subreddits. Before quarantine, the daily average of



(a) Participation to distinct subreddits (quarantine).

(b) Participation to distinct subreddits (restriction).

Fig. 12. ITS regressions of core users participation (either by commenting or by submitting content) to distinct subreddits around the quarantine and the restriction.

distinct subreddits in which core users of r/The_Donald participated —as a whole— was of 132.3 by submissions and 1078.9 by comments. Upon quarantine, our BSTS analysis indicates a significant (pp = .001) decrease of -8.6% by submissions and -11% by comments, albeit both immediate effects and trends are moderately downward, as visible from the ITS regression in Figure 12a. Upon restriction, however, the post-intervention downward trend and significant (pp = .001) relative effects are more pronounced, especially in the case of submissions at -33% with respect to comments at -11%, as shown in Figure 12b. Comparing results from Figure 12a with those of Figures 5 and 8 also allows gaining insights into the tactics —and their efficacy— used by core users of r/The_Donald to circumvent Reddit's interventions. After quarantine, core users posted many more submissions outside of r/The_Donald, as a way to elude the intervention. This behavior resulted in a markedly different submissions proclivity. However, the new submissions did not attract much attention since they received few comments, as testified by a relatively stable comment-based proclivity.

6 DISCUSSION

Our results allow to evaluate the effectiveness of quarantining and restricting r/The_Donald, both within the subreddit as well as in other parts of Reddit. In summary, the quarantine had mild positive effects with respect to reducing user activity and strong positive immediate effects for reducing toxicity within r/The_Donald. These effects were stronger for core users of r/The_Donald with respect to the other users. On the contrary however, quarantining r/The_Donald had no significant effects on news quality and, more worryingly, had the strong negative long-term effect of increasing toxicity. In particular, the degree of toxicity within r/The_Donald reverted to, and even surpassed, pre-intervention levels around 6 months after the quarantine.

Regarding the effects that quarantining and restricting r/The_Donald had on other subreddits, we found positive effects towards reducing user activity. Specifically, while the quarantine produced mild effects, the restriction was instead very effective at reducing activity. However, both the quarantine and restriction caused an overall increase in toxicity. In detail, the quarantine caused a small immediate increase and strong long-term decrease in toxicity, while the restriction caused moderate immediate increase as well as strong long-term increase in toxicity. Moreover, both interventions also had marked negative effects on the quality of the shared news, which gradually became more polarized and less factual.

Overall, our results support the following conclusions:

- The restriction produced stronger effects platform-wise than the quarantine.

- Core users suffered stronger effects than other users.
- Both interventions did well at reducing activity.
- At the same time however, both interventions produced an increase in toxicity.
- Both also caused affected users to share more polarized and less factual news.

In the next paragraphs we discuss our results in the context of the current state of knowledge, the main limitations of our work, and we identify possible directions for future research.

6.1 Temporal Variations of Effects

Longitudinal studies on moderation intervention effects are almost absent in literature, with most existing studies making no distinction between short-term and long-term effects. Exceptions are Chandrasekharan et al. who raised attention to the importance of verifying «whether the subreddits that come out of quarantine maintain improved discourse over long periods of time or whether they return to incendiary behavior», although without providing results [10]. Instead, Jhaver et al. evaluated the effects of deplatforming influential Twitter users and claimed that this had long-term positive consequences [29]. In this regard our results suggest caution, as we found important differences between the short-term and long-term effects of quarantine and restriction. Such differences were striking for toxicity within `r/The_Donald` after the quarantine. In that case, the strong initial decrease in toxicity started decaying a few days after the intervention, up to the point that 6 months later, toxicity had surpassed pre-intervention levels.

These results relate to the existing literature in a number of ways. Firstly, Chandrasekharan et al. pointed that particular attention should be devoted to monitoring toxic behavior when quarantines are lifted [10], implying that removing an intervention might revert the positive effect that it originally had. Here however, we observed a progressive reversal of the effect just a few days after the quarantine was enforced (not lifted). Our results thus go beyond Chandrasekharan et al.'s concerns, suggesting that intervention effects should be continuously evaluated through time, not only in the aftermath of major changes. In addition, our results about the temporal variations of intervention effects also suggest that more attention should be paid to this aspect. Therefore, in future studies it would be interesting to separately evaluate short- and long-term effects. Moreover, results of previous studies that do not make this distinction should be taken with care.

Our results also provide useful guidance for policing online platforms, in that platform administrators should continuously monitor the effects of their interventions and, in case, should reiterate them when the initial advantages have decayed. Some interventions should thus not only be seen as a one-off medicine, but rather as continuous or recurring treatments, depending on the need.

6.2 Spillover Effects

When a community suffers a moderation intervention, part of its members migrate to other communities or platforms. This long-studied behavior is inevitable when a community is permanently closed (e.g., deplatformed) [11, 28, 32, 37]. However, it has been shown that even milder interventions might induce some users to migrate [10]. When evaluating interventions it is thus important to assess changes not only within the moderated community, but also in other communities to which affected users might have migrated. We call the latter type of changes *spillover effects*.

Chandrasekharan et al. saw such effects in the quarantines of `r/The_Donald` and `r/TheRedPill`, finding that the intervention decreased the activity levels and influx of new users in many other communities frequented by users of `r/The_Donald` [10]. In other words, not only did the quarantine reduce activity and participation within the moderated subreddit, but it also had a similar effect in related subreddits. Our analysis confirms these results. Then, based on their findings about user activity, Chandrasekharan et al. concluded that «quarantining `r/The_Donald` did not spread the infection to other parts of Reddit» [10]. Our results on toxicity and quality of shared news provide additional evidence to evaluate this claim. In detail, we found that the quarantine caused a small immediate surge and a strong long-term reduction in the toxicity of core users of r/The_Donald when they commented in other subreddits. However, it also caused those users to share much more low-quality (i.e., politically biased and less factual) information. Overall, our results paint a more complex and nuanced picture of the effects that quarantining r/The_Donald had on other parts of Reddit. In conclusion, our results call for additional efforts at assessing spillover effects of moderation interventions, especially for milder interventions compared to deplatforming.

6.3 Side Effects and Nuanced Intervention Effects

Our work, as well as most existing studies on the effectiveness of Reddit's moderation interventions, investigated quarantines, restrictions, and bans [10, 11, 28, 45]. All three have different goals: quarantines deter interaction, restrictions limit content creation, and bans permanently shut subreddits. Still, it is worth noting that while their mechanics directly affect user activity and participation, they do not directly impact other aspects, such as the habits of the moderated community or the way in which the users express themselves. As such, it is safe to assume that all three aforementioned interventions were primarily designed to reduce user activity in problematic subreddits. For this specific objective (i.e., reducing activity), all existing studies support the efficacy of Reddit's interventions [10, 11, 28, 45] and our new results strongly confirm these findings.

However, a single intervention is capable of affecting multiple dimensions of user behavior [33], and even interventions designed with a specific objective (e.g., reducing activity) might cause a number of side effects. Our results provide evidence that despite achieving the desired objective of reducing the activity of problematic users within and outside of r/The_Donald, the sequence of interventions also had the *undesired side effects* of increasing the toxicity of such problematic users and of reducing the quality of the news shared and consumed by them (i.e., shared articles became more polarized and less factual). In previous literature, we found mixed support for our findings and many contrasting results. Chandrasekharan et al. evaluated the effects of banning two hateful subreddits, concluding that former members greatly reduced their hate speech and did not cause an increase in hate speech in the subreddits to which they migrated afterward [11]. In a subsequent study on two quarantines [10], some of the authors from the previous study obtained different results, finding that users in quarantined subreddits did not change their use of toxic terms. They concluded that the quarantine did not serve the goal of reducing offensive posts in the moderated community since it left this dimension essentially unaffected. Horta Ribeiro et al. reached yet another conclusion, finding a rise in toxicity following the restriction of two subreddits [28].

Our results are aligned with findings by Horta Ribeiro et al. [28], since we both measured a significant surge in toxicity following Reddit's interventions. At the same time however, ours also question their interpretation of these findings. Horta Ribeiro et al. explained their results in terms of the affordances of social platforms, concluding that the rise in toxicity observed when users migrated to fringe and unregulated platforms could be the «consequence of the removal of platform moderation» [28]. However, in our study we measured a similar rise in toxicity without any reduction in platform moderation, since we studied users who remained on the platform. If anything, our users even faced stronger moderation as they migrated to other subreddits with stricter content policies and rules. An alternative explanation was proposed by Chandrasekharan et al. in the context of Reddit's quarantines [10]. They postulated that quarantines could increase the insularity [1] (i.e., polarization and radicalization) of moderated communities and push users toward more extreme positions. Our results seem to confirm this interpretation, since we measured increases in both polarization and toxicity after Reddit quarantined and restricted r/The_Donald.

The above discussion provides theoretical contributions to the understanding of the effects of moderation interventions. In addition to these, our results also suggest some practical research and policing guidelines. The presence of a multitude of consequences following a moderation intervention mandates care in drawing conclusions on an intervention's efficacy. In particular, future studies should carry out nuanced analyses to assess effects across many behavioral dimensions. In the context of health interventions —a field from which the study of online moderation interventions inherits many characteristics— patients that experience side effects of a treatment can provide feedback to their physicians. Here, we lack this valuable feedback and must therefore pay extraordinary attention to the possible unintended consequences of moderation interventions.

6.4 Implications of Reducing User Activity

To date, all studies that evaluated the effects of Reddit's interventions found evidence for their effectiveness at *reducing the activity* of problematic users. Hence, many studies concluded that such interventions had largely positive effects [11, 28, 29, 45]. However, their effectiveness depends on the objective that Reddit administrators had when they enforced them, which raises a conundrum. On the one hand, Reddit administrators stated that «one of the primary goals of quarantining is to compel users to rethink their behavior and reduce offensive posts» [10], which suggests that their objective was related to *reducing the toxicity* of problematic users. On the other hand however, we noted (§6.3) that the mechanics of Reddit's interventions are such that they only directly affect user activity, rather than their toxicity. This reflection surfaces a discrepancy between the motivations stated by Reddit administrators and the moderation interventions that they deployed. Interestingly, the existence of this conceptual discrepancy explains the practical results of our study, where we measured a strong reduction in activity, counterbalanced by an overall increase in toxicity and in the sharing of biased news. More importantly, the discrepancy also suggests that current community-level interventions such as quarantines, restrictions, and bans might be *misdirected* and hence ineffective, or at the very least inefficient, at supporting platform's objectives.

In addition to the above misdirection issue, Reddit's community-level interventions might cause a second problem. Indeed, interventions that reduce activity cause user migrations and, depending on the scope and severity of the intervention, a large number of users might decide to migrate to alternative platforms [28]. Platforms might therefore be disincentivized to apply such interventions, for fear of losing users and revenues [8]. As the only "solution" to this issue, current literature appealed to platform's ethical principles, underlining the importance of pursuing the goal of platform moderation even at the cost of reduced revenues [29].

The two aforementioned problems highlight some of the limitations of community-level interventions, which appear to be unfit and inappropriate for the current moderation needs of online platforms. For the future, it would thus be advisable to design and deploy nuanced interventions that are more in focus with the objectives of online moderation. To this end, the recent development and experimentation with soft moderation interventions [54], as opposed to hard interventions like deplatforming, could be considered as a promising initial step in the right direction. Then, our findings suggest other possible directions of research, in addition to soft interventions. For instance, the fact that different (groups of) users suffered, in general, different effects to the same intervention —as in the case of core users of `r/The_Donald`— could motivate the deployment of *diversified* and *personalized* interventions [16]. So far, online moderation has always followed a "one-size-fits-all" approach, where each intervention was applied in the same way for all users, neglecting individual differences and the advantages of personalization. Instead, for the future we envision the possibility to deploy personalized interventions, each specifically designed for and targeted to a different group of users [16]. This novel approach to online moderation will likely boost the effectiveness of moderation interventions, by emphasizing the desired effects of the moderation, while at the same time minimizing the possible undesired side effects. Designing nuanced interventions also relates to Kiesler et al.'s theory of graduated sanctions [31]. These could contribute not only at making platforms more accountable for their moderation [10], but also at having more targeted interventions, capable of reducing negative side effects such as user migration and loss of revenues. The design and evaluation of targeted and nuanced interventions thus represents fertile ground for future research and experimentation on online moderation.

6.5 The Current Context for Platform Interventions

Since the 2016 Donald Trump presidential win and the unexpected outcome of the UK Brexit referendum, social media platforms have been facing a tremendous pressure to take action against issues such as misinformation and online misbehavior. Recently, the pressure heightened even more as a consequence of other dramatic events such as the George Floyd protests and the emergence of the COVID-19 infodemic. As a result of those events the platforms responded to the growing pressure by hastily enforcing a number of interventions. As notable examples within the scope of our study, Reddit issued several changes to its policies, including those addressing the George Floyd protests murder[7] that eventually led to the ban of `r/The_Donald`.[8]

Despite appearing as reasonable solutions and serving as public evidence of the platforms' willingness to tackle the issues they contributed to create, many of such interventions were devised and applied light-mindedly. Post-hoc scientific analyses revealed that many interventions proposed in recent years produced mixed effects or no effects at all [9, 10, 12, 28], and some even exacerbated the problems they were trying to solve [2, 21, 40]. Pennycook and Rand concluded a 2020 op-ed on The New York Times[9] remarking that moderation interventions should «not just rely on common sense or intuition» but that should instead be «empirically grounded». In consideration of our mixed results on the effectiveness of the sequence of interventions on `r/The_Donald`, we reiterate this recommendation to balance the timeliness of an intervention with its empirical soundness, motivating present and future research along this important direction.

6.6 Limitations and Future Work

6.6.1 Selection Bias. On the one hand, our choice of focusing on the sequence of moderation interventions that targeted `r/The_Donald` allowed us to thoroughly analyze the former most prominent political community on Reddit. Furthermore, it allowed us to compare and discuss our results with those of several other existing studies that were so far disconnected. For example, we compared our results about the effectiveness of the quarantine and the restriction with the results from [10] and [28], respectively. We also compared our results with those of previous works that studied other deplatforming interventions [11, 29, 45]. Our work thus contributed to reconcile many of the existing studies on the topic. On the other hand, our analyses are solely focused on the moderation interventions enforced to `r/The_Donald` and, as such, might suffer from selection bias and lack of generality. Therefore, more research is needed in order to verify whether our findings hold also for other communities of users, which could react differently to Reddit's interventions. To this end, our study provides useful guidelines for future research aimed at assessing more precisely the consequences of online moderation strategies.

6.6.2 Confounders and Exogenous Causes. An inherent limitation when estimating causal effects with observational data is the susceptibility of the measured quantities to possible exogenous causes (i.e., confounders). Our study, as well as many others [10, 11, 28, 29], is affected by this limitation. In

[7] https://redd.it/gxas21

[8] https://redd.it/hi3oht

[9] https://www.nytimes.com/2020/03/24/opinion/fake-news-social-media.html

fact, our analysis did not explicitly consider possible exogenous events that might have contributed to cause some of the effects that we measured. The most noteworthy of such events is the murder of George Floyd, occurred on May 25, 2020, which lays within our post-restriction time window. As such, results about the effects of the restriction might be influenced by this event, as visible in Figure 9 and as already noted in [28]. To circumvent this limitation, we repeated the BSTS analysis about toxicity changes caused by the restriction, by only considering data up to the killing of George Floyd. The results of this additional analysis confirm our initial findings, in that the restriction caused a marked increase in toxicity. In other words, our conclusions are robust to this exogenous event. Nevertheless, for the future it is advisable to adopt causal inference methodologies that allow to account for, at least, straightforward and known exogenous causes. To this regard, the BSTS method that we adopted to provide our quantitative results admits the possibility to model known exogenous events [7], and thus represents a favorable methodology for future causal analyses on the effects of online moderation interventions.

6.6.3 Multidimensionality of Effects. When elaborating on the possible consequences of the exposure to fake news, Lazer et al. concluded that they might have «many potential pathways of influence, from increasing cynicism and apathy to encouraging extremism», in addition to the obvious consequence of affecting political preferences [33]. In other words, fake news might have *many and diverse effects* spread across multiple dimensions of user behavior and ideology. The same can be said about the interventions that the platforms put in place to contrast fake news and the other ailments that affect online social spaces. Driven by this consideration, in this work we carried out a nuanced analysis and we moved beyond existing studies that almost solely investigated changes in activity and toxicity (or hate speech) [10, 11, 28, 29, 45]. We did this by including several important dimensions and metrics that were so far overlooked. Among them is the analysis of the quality of shared news articles, which in turn provides insights into the degree of political polarization and factual reporting of the news consumed by the analyzed communities, and the group community proclivity. Nonetheless, the interventions investigated in our work might have affected many additional dimensions of user behavior and ideology. For example, there might have been changes in user stances toward certain relevant (e.g., political) topics, in their trust of authoritative institutions, and in other emotional, social, and psychological dimensions. Because of this, our results likely provide only a partial view of the full extent of effects caused by the interventions issued on `r/The_Donald`. For the future it is thus important to progressively make measurable a larger set of dimensions and to investigate intervention effects also in these dimensions.

6.6.4 User-level Effects. Our study considers effects aggregated at the community-level, like most of the works on the same subject [10, 11, 28, 29]. Indeed, aggregated community-level effects are perhaps the most natural and direct way to evaluate community-level interventions (e.g., ban of an entire community). Nonetheless, a community-level post-intervention effect is the combination of many and potentially diverse user-level effects. Hence, the aggregated community-level effect might be weakly representative of the underlying behavior of individuals or smaller user groups. For example, Saleem and Ruths measured user-level changes in comment activity and subreddit participation that resulted from the ban of two problematic communities on Reddit [45], showing very diverse user-level effects. Based on these considerations, it would be interesting to evaluate user-level effects for a larger set of moderation interventions. In particular, for future work we aim at assessing whether certain community-level effects are the result of homogeneous or heterogeneous user behavior. In the latter case, it would be interesting to investigate which user-level characteristics determine effect diversity, and thus study the possibility of preemptively identifying users likely to significantly deviate from intervention expectations. In either case, understanding effects at the user-level might increase our chances to design and deploy more effective interventions.

7 CONCLUSIONS

We carried out a multidimensional causal analysis of the sequence of moderation interventions enforced on `r/The_Donald`. Our results paint a nuanced picture of the effects of such interventions and support the following take-away messages: (i) the restriction produced stronger effects platform-wise than the quarantine, (ii) core users of `r/The_Donald` suffered stronger effects than other users, (iii) both the quarantine and the restriction significantly reduced user activity, however (iv) both also caused an increase in toxicity and (v) caused users to share more polarized and less factual news. We conclude that the sequence of interventions had mixed effects. For the future, it will be important to advance the understanding and the development of moderation interventions, so as to obtain tools capable of achieving the objectives of online moderation with minimal side effects.

REFERENCES

[1] Kimberley Allison and Kay Bussey. 2020. Communal quirks and circlejerks: A taxonomy of processes contributing to insularity in online communities. In *The 14th International AAAI Conference on Web and Social Media (ICWSM'20).*

[2] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *The 14th International AAAI Conference on Web and Social Media (ICWSM'20).* 830–839.

[4] Yochai Benkler, Robert Faris, and Hal Roberts. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics.* Oxford University Press.

[5] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International Journal of Epidemiology* 46, 1 (2017), 348–355.

[6] Alexandre Bovet and Hernán A Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications* 10, 1 (2019), 1–14.

[7] Kay H Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. 2015. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* 9, 1 (2015), 247–274.

[8] Caitlin Ring Carlson and Luc S Cousineau. 2020. Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's Quarantine Practice. *Journal of Media Ethics* 35, 4 (2020), 202–213.

[9] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *The 19th ACM Conference on Computer-supported Cooperative Work And Social Computing (CSCW'16).* 1201–1213.

[10] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2020. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *arXiv:2009.11483* (2020).

[11] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. In *The 20th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW'17).* 1–22.

[12] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *The 8th International AAAI Conference on Web and Social Media (ICWSM'14).*

[13] Simon Copland. 2020. Reddit quarantined: Can changing platform affordances reduce hateful material online? *Internet Policy Review* 9, 4 (2020), 1–26.

[14] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design frictions for mindful interactions: The case for microboundaries. In *The 34th CHI Conference on Human Factors in Computing Systems (CHI'16).* 1389–1397.

[15] Stefano Cresci. 2020. A decade of social bot detection. *Commun. ACM* 63, 10 (2020), 72–83.

[16] Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In *The 33rd ACM Conference on Hypertext and Social Media (HT'22).* ACM, 248–251.

[17] Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *The 29th International Joint Conference on Artificial Intelligence (IJCAI'20).* 4826–4832.

[18] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *The 11th International AAAI Conference on Web and Social Media (ICWSM'17).*

[19] Hongbo Deng, Michael R Lyu, and Irwin King. 2009. A generalized Co-HITS algorithm and its application to bipartite graphs. In *The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09).* 239–248.

[20] Roberto Di Pietro, Maurantonio Caprolu, Simone Raponi, and Stefano Cresci. 2021. *New Dimensions of Information Warfare*. Advances in Information Security, Vol. 84. Springer. 268 pages.

[21] Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).

[22] Claudia Flores-Saviaga, Brian Keegan, and Saiph Savage. 2018. Mobilizing the Trump train: Understanding collective action in a political trolling community. In *The 12th International AAAI Conference on Web and Social Media (ICWSM'18)*.

[23] Paula Fortuna, Juan Soler-Company, and Leo Wanner. 2021. How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management* 58, 3 (2021), 102524.

[24] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *The 12th International AAAI Conference on Web and Social Media (ICWSM'18)*.

[25] Tiana Gaudette, Ryan Scrivens, Garth Davies, and Richard Frank. 2020. Upvoting extremism: Collective identity formation and the extreme right on Reddit. *New Media & Society* (2020), 1461444820958123.

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

[27] Christoph Hanck, Martin Arnold, Alexander Gerber, and Martin Schmelzer. 2019. Introduction to Econometrics with R. *University of Duisburg-Essen* (2019), 1–9.

[28] 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. In *The 24th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW'21)*. 1–24.

[29] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. In *The 24th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW'21)*. 1–30.

[30] Andreas Jungherr, Oliver Posegga, and Jisun An. 2021. Populist supporters on Reddit: A comparison of content and behavioral patterns within publics of supporters of Donald Trump and Hillary Clinton. *Social Science Computer Review* (2021), 0894439321996130.

[31] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building successful online communities: Evidence-based social design* (2012), 125–178.

[32] Shamanth Kumar, Reza Zafarani, and Huan Liu. 2011. Understanding user migration patterns in social media. In *The 25th AAAI Conference on Artificial Intelligence (AAAI'11)*.

[33] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[34] Joan Massachs, Corrado Monti, Gianmarco De Francisci Morales, and Francesco Bonchi. 2020. Roots of Trumpism: Homophily and social feedback in Donald Trump support on Reddit. In *The 12th ACM Conference on Web Science (WebSci'20)*. 49–58.

[35] Marcelo Mendoza, Maurizio Tesconi, and Stefano Cresci. 2020. Bots in social and interaction networks: Detection and impact estimation. *ACM Transactions on Information Systems* 39, 1 (2020), 1–32.

[36] Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2020. "And We Will Fight for Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. In *The 14th International AAAI Conference on Web and Social Media (ICWSM'20)*, Vol. 14. 452–463.

[37] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on Reddit during a period of community unrest. In *The 10th International AAAI Conference on Web and Social Media (ICWSM'16)*.

[38] Leonardo Nizzoli, Serena Tardelli, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. 2021. Coordinated behavior on social media in 2019 UK General Election. In *The 15th International AAAI Conference on Web and Social Media (ICWSM'21)*. 443–454.

[39] Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Soros, Child Sacrifices, and 5G: Understanding the Spread of Conspiracy Theories on Web Communities. *arXiv:2111.02187* (2021).

[40] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66, 11 (2020), 4944–4957.

[41] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *The 14th International AAAI Conference on Web and Social Media (ICWSM'20)*, Vol. 14. 557–568.

[42] Tyler T Reny and Benjamin J Newman. 2021. The opinion-mobilizing effect of social protest against police violence: Evidence from the 2020 George Floyd protests. *American Political Science Review* (2021), 1–9.

[43] Bernhard Rieder and Yarden Skop. 2021. The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data & Society* 8, 2 (2021), 20539517211046181.

[44] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 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 (2021), 20563051211052906.

[45] Haji Mohammad Saleem and Derek Ruths. 2018. The aftermath of disbanding an online hateful community. *arXiv:1804.07354* (2018).

[46] Andrea L Schaffer, Timothy A Dobbins, and Sallie-Anne Pearson. 2021. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. *BMC Medical Research Methodology* 21, 1 (2021), 1–12.

[47] Steven L Scott and Hal R Varian. 2014. Predicting the present with Bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation* 5, 1-2 (2014), 4–23.

[48] Ryan P Shepherd. 2020. Gaming Reddit's algorithm: r/The_Donald, amplification, and the rhetoric of sorting. *Computers and Composition* 56 (2020), 102572.

[49] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. 2019. A characterization of political communities on Reddit. In *The 30th ACM Conference on Hypertext and Social Media (HT'19)*. 259–263.

[50] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe* 27 (2017).

[51] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* 28, 4 (2010), 1–38.

[52] Galen Weld, Maria Glenski, and Tim Althoff. 2021. Political Bias and Factualness in News Sharing across more than 100,000 Online Communities. In *The 15th International AAAI Conference on Web and Social Media (ICWSM'21)*.

[53] Tim Weninger. 2014. An exploration of submissions and discussions in social news: Mining collective intelligence of Reddit. *Social Network Analysis and Mining* 4, 1 (2014), 173.

[54] Savvas Zannettou. 2021. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. In *The 15th International AAAI Conference on Web and Social Media (ICWSM'21)*. 865–876.

[55] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2017. The Web centipede: Understanding how Web communities influence each other through the lens of mainstream and alternative news sources. In *The 2017 Internet Measurement Conference (IMC'17)*. 405–417.

[56] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who let the trolls out? Towards understanding state-sponsored trolls. In *The 11th ACM Conference on Web Science (WebSci'19)*. 353–362.

Received January 2022; revised April 2022; accepted August 2022