Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter

ANONYMOUS AUTHOR(S)

Deplatforming refers to the permanent ban of controversial public figures with large followings on social media sites. In recent years, platforms like Facebook, Twitter and YouTube have deplatformed many offensive influencers to curb the spread of offensive speech. We present a case study of three high-profile influencers who were deplatformed on Twitter—Alex Jones, Milo Yiannopoulos, and Owen Benjamin. Working with over 49M tweets, we found that deplatforming significantly reduced the number of conversations about all three influencers on Twitter. Further, analyzing the Twitter-wide activity of these influencers' supporters, we show that the overall activity and toxicity levels of supporters declined after deplatforming. We contribute a methodological framework to systematically examine the effectiveness of moderation interventions and discuss broader implications of using deplatforming as a moderation strategy.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation; freedom of speech; platform governance

ACM Reference Format:
Anonymous Author(s). 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *J. ACM* 37, 4, Article 111 (February 2021), 28 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Social media platforms play an increasingly important civic role as spaces for individual self-expression and collective association. However, the freedom these platforms provide to their content creators also gives individuals who promote offensive speech[1] an opportunity to amass followers. For instance, Milo Yiannopoulos, a British far-right political commentator, gathered a large army of anonymous online activists on Twitter who amplified his calls for targeted harassment [46]. Another extremist influencer, Alex Jones, marshalled thousands of followers on social media to promote his conspiracy theories, which led to violent acts [66]. One way to contain the spread of offensive speech online is to identify key influencers who lead communities that promote such speech and *deplatform* them, or, in other words, permanently ban them from the platform.

Removing someone from a platform is an extreme step that should not be taken lightly. However, platforms have rules for appropriate behavior, and when a site member breaks those rules repeatedly, the platform may need to take action. The toxicity created by influencers who promote offensive speech and their supporters can also impact vulnerable user groups, making it crucial for platforms to attend to such influencers' activities. We do not attempt to address the question of exactly when deplatforming is advisable. Rather, we address the underlying empirical question: *what*

[1]For the purposes of this paper, we use the term 'offensive' to mean promoting toxic speech. This includes sexist, racist, homophobic, or transphobic posts and targeted harassment as well as conspiracy theories that target certain racial or political groups.

0004-5411/2021/2-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

happens when someone is deplatformed? Understanding what happens after an influencer has been deplatformed may help social media providers make sound decisions about whether and when to employ this technique.

Previous work has addressed what happens when toxic online *groups* get moderated. For example, Chandrasekharan et al. [17] and Saleem and Ruths [81] independently analyzed Reddit's bans of racist and fat-shaming communities in 2015; they determined that the "ban worked": community members who remained on the platform dramatically decreased their hate speech usage. Another study found that quarantining offensive communities on Reddit, i.e., impeding direct access to them, made it more difficult for them to recruit new members [16].[2] While these studies focused on groups, our work focuses on the moderation of offensive *individuals* [19, 87] in a different setting: we examine the long-term consequences of deplatforming prominent influencers on Twitter.

As a moderation strategy [37], deplatforming has recently been on the rise [77]. Facebook, Twitter, Instagram, YouTube and other platforms have all banned controversial influencers for spreading misinformation, conducting harassment, or violating other platform policies [11, 25, 56, 57]. Mainstream news outlets have extensively covered different instances of deplatforming and generally praised platforms for them [11, 56, 57]. However, it is unclear how deplatforming as a tactic affects the spread of influencers' ideas or the activities of their supporters *in the long run*. From the perspective of platforms, it should be relatively straightforward to identify accounts with thousands of followers who promote offensive content and deplatform those accounts. Therefore, it is vital that we examine whether this low-cost moderation intervention is a viable means to detoxify mainstream social media.

1.1 Deplatforming on Twitter

Today, Twitter is one of the most popular social media platforms. On Twitter, users can *follow* any account in order to receive the posts made by it, called *tweets*, in their news-feed. This subscription-without-permission model makes Twitter an ideal platform for influencers to gain huge audiences and promote their ideas. Many far-right influencers who promote offensive speech, like Alex Jones and Richard Spencer [98], gained widespread popularity because of their activities on Twitter.

When an influencer is deplatformed, all their tweets are removed, although replies to those tweets remain online. Deplatformed influencers can no longer create a new account under their real names, and any attempts to do so quickly result in bans. As a result, they are forced to simply leave the platform.

To date, we have only anecdotal evidence that deplatforming is an effective means to curb the spread of offensive behavior [68]. To contribute empirical evidence that speaks to this issue, we study the effects of deplatforming three promoters of offensive speech on Twitter: (1) Alex Jones, an American radio show host and political extremist who gained notoriety for promoting many conspiracy theories, (2) Milo Yiannopoulos, a British political commentator who used his celebrity to incite targeted harassment and became known for ridiculing Islam, feminism and social justice, and (3) Owen Benjamin, an American alt-right actor, comedian and political commentator who promoted many anti-Semitic conspiracy theories, including that the Holocaust was exaggerated, and held anti-LGBT views. Through studying Twitter activity patterns around these influencers and their thousands of supporters, we offer initial insights into the long-term effects of using deplatforming as a moderation strategy.

[2] *Banning* a Reddit community removes all its prior content and blocks access to it, while *quarantining* a community stops its content from appearing in search results or on the Reddit front page. While banning and quarantining of subreddits are both examples of community-level moderation interventions, *deplatforming* is a user-level intervention that refers to permanent bans of individual accounts.

1.2 Research Questions

Some critics predict that deplatforming might not work because banning people may draw attention to them or their ideas that platforms are trying to suppress [24, 68]. This idea is often referred to as the *Streisand effect*[3], and many examples of this effect have previously been observed [45, 60]. To evaluate whether the Streisand effect occurred in the context of deplatforming, we ask:

> **RQ1**: How did deplatforming affect the number of conversations about deplatformed influencers?

Critics argue that "any kind of censorship can create an aura of conspiracy that makes forbidden ideas attractive" [24]. To test whether deplatforming renders the ideas spread by influencers more popular on Twitter, we ask the following research question:

> **RQ2**: How did deplatforming affect the spread of offensive ideas held by deplatformed influencers?

Those suspicious of the effect of deplatforming also argue that even when banning someone reduces his or her audience, it can, at the same time, strengthen the audience that remains [24, 68]. They suggest that deplatforming could ignite supporters who view it as a free speech violation. Indeed, deplatforming has been used to express victimization and gather support for sanctioned celebrities [77]. Others argue that deplatforming would effectively drive extremist supporters of deplatformed influencers to other spaces and reduce their negative impact on the platform. To this end, we focus on the most ardent supporters of influencers and ask:

> **RQ3**: How did deplatforming affect the overall activities of supporters of these deplatformed influencers?

1.3 Summary of Methods, Findings and Implications

Methods. We answer our research questions by examining observational data from Twitter through a temporal analysis of (1) tweets directly referencing deplatformed influencers, (2) tweets referencing their offensive ideas, and (3) all tweets posted by their supporters. For each influencer, we limit our data collection and analyses to the tweets posted in the period six months before and after their deplatforming. Working with over 49M tweets, we chose metrics [103] that include posting volume and content toxicity scores obtained via the Perspective API. We then used *causal inference methods* (specifically, interrupted time series regression analyses) and Wilcoxon signed-rank tests to examine variations in levels of activity and toxicity, accounting for ongoing temporal trends.

Findings. We determined that deplatforming disrupted the discussions about influencers—posts referencing each influencer declined significantly, by 91.77% on average. Additionally, the number of unique users and new users tweeting about each influencer also diminished significantly, by 89.51% and 89.65% on average, respectively. Further, *deplatforming significantly reduced the popularity of many anti-social ideas* associated with influencers. We also found that *deplatforming significantly reduced the overall posting activity levels of supporters* in each case: median drop in the volume of tweets posted by supporters averaged 12.59% across the three influencers. Finally, *deplatforming significantly reduced the overall toxicity levels of supporters of each influencer*: across the three cases, the median decline in toxicity score of tweets posted by supporters averaged 5.84%.

[3]The term "Streisand Effect" emerged from an incident where the actress Barbra Streisand attempted to restrict online views of her Malibu mansion on a public website by filing a lawsuit against a photographer. However, the publicity created by this photograph had the paradoxical effect of stimulating public interest and resulting in many more views than if she had done nothing [45].

Implications. Our work demonstrates the efficacy of deplatforming offensive influencers to counteract offensive speech in online communities. We show that deplatforming can lead to second-order harms and offer methodological insights into how platforms can defend against them. We provide a computational framework that internet platforms can use to evaluate the longitudinal knock-on effects of a wide range of moderation interventions. Our findings offer empirical evidence to support calls for more proactive monitoring and sanctioning of disruptive influencers.

2 BACKGROUND AND RELATED WORK

We next provide background information on content moderation, deplatforming, extremist actors and toxicity in online communities, and review key related work.

2.1 Content Moderation and Deplatforming

Content moderation refers to the sociotechnical mechanisms that platforms deploy to screen user-generated content and determine its appropriateness [75]. Many platforms employ a large group of commercial content moderators or freelancers who work with them on contract to detect misbehaviors at massive levels of scale [76]. Other platforms, like Reddit, Discord and Twitch, rely on end-users to serve as volunteer moderators [48, 51, 59, 65, 97]. These moderators are assisted by automated tools that remove inappropriate posts or flag them to be reviewed by moderators [48].

In recent years, scholars and mainstream news outlets have frequently criticized social media platforms for not doing enough to efficiently moderate their content [3, 23, 69, 95]. In response, platforms have stepped up their content moderation efforts to curb the online spread of hate speech, misinformation and conspiracy theories. For example, Twitter has labeled tweets that violate its guidelines, including those posted by the US president Trump, as offensive [80]. Pinterest has blocked search results for anti-vaccination queries in an effort to fight the spread of vaccine misinformation [96]. Reddit has banned certain toxic communities and quarantined others [16, 17]. Facebook has identified Nazi and white nationalist groups on its platform and shared them with nonprofits who help people leave such hate groups [25]. We argue that deplatforming is another important tool in the arsenal of moderation interventions available to platforms. In this work, we contribute an understanding of what happens when platforms use this tool.

Hate speech is illegal in much of the world. It remains legal in the United States; however, private platforms may still remove it if they wish. US laws about freedom of speech control what the government can do — not what private entities can do [48]. Social media platforms are protected by law when they take actions like deplatforming. Section 230 of the U.S. Communications Decency Act grants immunity to internet intermediaries to regulate user-generated content however they see fit [30]. Additionally, this law does not consider platforms responsible for upholding their users' freedom of speech [35, 55].

While much of the moderation work on social media platforms involves removing posts that do not comply with their guidelines [34], moderation decisions like banning entire communities on Reddit or deplatforming influential public figures on Facebook or Twitter are more delicate. Such decisions can have long-ranging consequences; therefore, platforms may require a more nuanced analysis of the potential benefits and harms of those decisions. These decisions also have financial implications. Banning an influencer who encourages offensive speech may curb the spread of toxic ideologies on the platform, but it may also push their supporters away from the platform. To our knowledge, our work demonstrates the first empirical evidence of the effectiveness of deplatforming in reducing toxicity online. Such evidence can support platforms in making wise decisions about whether and when to use this technique.

2.2 New Media and Extremist Actors

Extremist actors frequently complain that mainstream news outlets are biased against them, do not sufficiently cover them, or misrepresent their views [63, 94]. Driven by this hostility towards the mainstream press, these extremists have actively turned to social media sites for mobilization, coordination, and information dissemination [83].

While the groups engaged in extremist media manipulation include a variety of participants (e.g., internet trolls, men's right activists and hyper-partisan news outlets), some of these participants become influencers who play a distinct role in media manipulation. These actors often use attention-hacking techniques [14] to gain prominence and then leverage their notoriety for increased coverage. Once they gain large followings, they use their talents to amplify fringe beliefs.

In the current work, we study three such extremist influencers, who gained widespread popularity on the social media site Twitter and then used that popularity to spread offensive speech. Next, we provide background on these three deplatformed influencers. We warn the reader that the next sections contain offensive ideas and language; however, we feel that this content is important to understanding the person being deplatformed.

2.2.1 Alex Jones. Alex Jones, the owner and primary operator of the website *Infowars*, has made a name for himself as a powerful force in the world of conspiracy theories and alternate news media. Many have criticized Jones and *Infowars* as manufacturers of damaging and often defamatory misinformation. Most famously, Jones perpetuated the myth that the 2012 mass shooting at Sandy Hook Elementary was a staged operation intended to advance gun control legislation [92]. In 2018, several families of Sandy Hook victims filed a law suit against Jones claiming that his false statements provoked continued threats and harassment from his followers [92]. Jones's unfounded claims also include suggesting a link between vaccines and autism [43], the existence of government-manufactured weapons to control weather [44], and more recently that a toothpaste sold in his supplement store "kills the whole SARS-corona family at point-blank range" [61].

In early August 2018, numerous sites such as Facebook, Apple, Youtube, Spotify, Vimeo, Pinterest, Mailchimp, and LinkedIn removed his account, with most citing the hateful nature of Jones's content as the reason for the intervention [88]. Twitter followed suit in September 2018 by banning Jones following his targeted harassment of CNN reporter Oliver Darcy [79].

2.2.2 Milo Yiannopoulos. Yiannopoulos, a self-described "supervillain" [1], established himself as a prominent figure in the far-right movement when writing for Breitbart News. An early supporter of the controversial GamerGate movement [64], Yiannopoulos fanned the flames of targeted harassment against several prominent women in the game industry through his news articles [100, 101] and social media posts. Yiannopoulos's Twitter account was banned in July 2016 following his targeted harassment of actress Leslie Jones [78].

Following his Twitter suspension, Yiannopoulos has continued to stir up controversy. Reports suggest he regularly corresponded with self-proclaimed white nationalists when formulating ideas for his Breitbart articles, sang "America the Beautiful" at a karaoke bar to a crowd of white supremacists giving Nazi salutes [10], and even lost his position at Breitbart after arguing that sexual relationships between young teenage boys and adults in some cases can "happen perfectly consensually" [71]. Yiannopoulos's actions have resulted in bans from several platforms beyond Twitter, including but not limited to Facebook, Instagram, and Patreon [6, 29].

2.2.3 Owen Benjamin. Benjamin, a comedian by trade, has used his influence to spread a wide variety of hate speech, misinformation and conspiracy theories. He is particularly known for posting unquestionably racist tweets. For instance, he once tweeted about his wish for the return of slavery just so racial justice activist Shaun King "can be forced to be a slave" under the guise of comedy

[8]. Benjamin was permanently banned from Twitter following a series of tweets about the sexual maturity of the Parkland shooting survivor and gun control activist David Hogg, a minor at the time [53]. Following his Twitter ban, Benjamin continued to make conspiratorial and anti-Semitic claims that resulted in suspensions on multiple platforms such as Youtube and Paypal [36, 72].

2.3 Toxicity in Online Communities

Online communities experience a wide variety of toxic behaviors, such as incivility [13], harassment [12, 22, 50], trolling [20] and cyberbullying [58]. In recent years, researchers have explored the mechanics of such online behaviors. For example, Massanari noted that Reddit's algorithmic politics and its policies about offensive content encourage "toxic technocultures" that normalize anti-feminist and misogynistic activism [64]. Kwak et al. studied toxicity in League of Legends, an online multiplayer game, and found that toxic behaviors like cyberbullying are often explained by attribution theory, i.e., toxic team players look for someone other than themselves to blame for poor team performance [58]. Zannettou et al. examined the dissemination of toxic memes across the web [102]. We add to this literature by investigating how deplatforming influencers affects the toxicity of their supporters on Twitter.

Though toxicity does not have a widely accepted definition, researchers have linked it to cyberbullying, profanity and hate speech [32, 62, 64, 70]. Given the widespread prevalence of toxicity online, researchers have developed multiple dictionaries and machine learning techniques to detect and remove toxic comments at scale [18, 32, 99]. Wulczyn et al., whose classifier we use (section 4.1.3), defined *toxicity* as having many elements of incivility but also a holistic assessment [99], and the production version of their classifier, Perspective API, has been used in many social media studies (e.g., [39, 41, 67, 73, 103]) to measure toxicity.

Prior research suggests that Perspective API sufficiently captures the hate speech and toxicity of content posted on social media [39, 41, 67, 73, 103]. For example, Rajadesingan et al. found that, for Reddit political communities, Perspective API's performance on detecting toxicity is similar to that of a human annotator [73], and Zanettou et al. [103], in their analysis of comments on news websites, found that Perspective's "Severe Toxicity" model outperforms other alternatives like HateSonar [26]. Some critics have shown that Perspective API has the potential for racial bias against speech by African Americans [21, 82], but we do not consider this source of bias to be relevant for our analyses because we use this API to compare the same individuals' toxicity before and after deplatforming.

3 DATA

In this section, we discuss data collection for our analysis, i.e., how we: selected deplatformed influencers for our case studies and collected tweets referencing these influencers; identified offensive ideas promoted by influencers and collected tweets mentioning these ideas; and identified influencers' supporters and collected their tweets.

3.1 Choosing Deplatformed Influencers

We selected influencer accounts for our case studies using the following criteria: (1) the account belongs to an individual instead of a company, organization or government entity and is associated with the individual's real identity, and (2) the account has at least 100,000 followers immediately before deplatforming.[4]

[4]Twitter's API does not provide historical follower counts. Therefore, we confirmed the number of followers of its deplatformed accounts by examining archived snapshots of their Twitter profiles on internet archives, e.g., archive.org's WayBack machine (https://archive.org/web/) and socialblade.com.

Table 1. Three selected deplatformed influencers, their deplatforming date, number of followers on that date, the number of tweets referencing influencers, the number of supporters we identified, and the number of tweets we collected for their supporters.

Influencer	# Followers	Deplatforming Date	# Tweets	# Supporters	# Supporters Tweets
Alex Jones	898,610	2018-09-06	1,157,713	2,935	17,050,653
Milo Yiannopoulos	338,000	2016-07-19	442,655	5,827	30,000,335
Owen Benjamin	122,634	2018-04-05	127,855	304	822,022

We began with a list of deplatformed influencers collected by researcher Richard Hanania [38]. Next, we augmented this seed list with additional influencer accounts by referring to news articles on Twitter deplatforming.[5] We filtered this list to include only those influencers that met our selection criteria. Through this process, we curated a list of 14 deplatformed influencers for our larger initiative on understanding different aspects of platform bans. For the current work, we focused on three deplatformed influencers as our case studies: Alex Jones, Milo Yionnapoulus, and Owen Benjamin. We selected these influencers because they had different levels of popularity, they all promoted toxic speech, and they were known for spreading different types of offensive ideas; moreover, their deplatforming was covered by multiple news outlets. We considered it important to select influencers who promoted toxic speech for this study because we were interested in examining how deplatforming affects the toxicity of their supporters. Therefore, although our augmented list contained influencers like Roger Stone and Martin Shkreli who are controversial public figures, we chose not to include them because they do not regularly engage in offensive speech. Table 1 lists the selected deplatformed influencers, their deplatforming date, and their follower count on that date.

3.2 Data for RQ1: Tweets Referencing Deplatformed Influencers

For each influencer, we collected only the tweets posted in the period six months before to six months after his deplatforming. We chose a threshold of six months to keep in line with other research on long-term effects of moderation interventions [17, 74, 91]. We also observed that the posting activity trends had stabilized at the ends of this time period and therefore we considered this an adequate time window to consider the long-term consequences. To meet our data collection needs, we developed a Twitter web scraper to collect all publicly accessible tweets.[6] We first identified a small set of Twitter hashtags and keywords that explicitly referenced each deplatformed account α and compiled them into a list. For this, we used Twitter's search feature and manually inspected the top results obtained through hand-curated search queries. For example, for Alex Jones, we used the search queries "Alex Jones" and "Infowars." Next, we used our web scraper to collect tweets containing any keyword in our list, and we called this collection α-D.[7]

Our inspection of α-D indicated that users generally referenced influencers either by using their name (e.g., "Alex Jones") or by using hashtags about them (e.g., "#FreeAlexJones"). Therefore, we added each influencer's name to the corresponding keywords list and further populated it using a snowball sampling approach: following the initial collection of α-D, we quantitatively inspected it for hashtags that most frequently co-occurred with items from the keywords list. Next, the first and second authors independently coded these hashtags to identify those that specifically referenced

[5]These articles were retrieved from sources like nytimes.com, cnet.com and thedailybeast.com. All accessed: 2019-11-10.
[6]Twitter allows its users to *protect* their tweets. Protected tweets are accessible only to followers who are manually approved by the account holder. Due to this restriction, our dataset contains only *public* tweets.
[7]Throughout our data collection, we conducted case-insensitive searches for tweets to capture the case-variant use of relevant keywords.

Table 2. Keywords used to collect tweets relevant to each deplatformed user. We used case-insensitive searches for these keywords during our data collection.

Deplatformed Influencer	Keywords Used to Curate Relevant Tweets
Alex Jones	alex jones, #freealexjones, #infowarsarmy, #iamalexjones, #realalexjones, #alex_jones, #infowarsban, #alexjonesbanned, #banalexjones, #unpluginfowars, #banalexjonesnow, #weareallalexjones, #thealexjoneschannel, #istandwithalexjones
Milo Yiannopoulos	milo yiannopoulos, #freemilo, #miloyiannopoulos, #milo, #freenero, #jesuismilo, #yiannopoulos, #milosd, #nero, #talktomilo, #freemilonow, #freethefaggot, #dangerousfaggot, #dangerousfaggottour, #milosexual, #miloatdepaul, #bannero, #jesuisnero, #freespeechformillo, #milobreakstheinternet, #bringbackmilo, #freemiloyiannopoulos, #fuckmilo, #suspendneroparty, #chokemilo, #milogirlsbreaktheinternet, #banmilo, #blockedformilo, #verifymilo, #miloatrutgers, #standwithmilo, #wheresmilo, #prayformilo, #teammilo
Owen Benjamin	owen benjamin, #owenbenjamin, #freeowenbenjamin, #freeowen, #imwithowen, #blameowen

Control Account	Handle	Corr. Influencer	# Followers	# Tweets
Drudge Report	drudge_report	Alex Jones	1,390,000	239,540
Steven Crowder	crowder	Milo Yiannopoulos	237,000	166,065
Gateway Pundit	gatewaypundit	Owen Benjamin	94,000	351,133

Table 3. Three influencers that have not been deplatformed and serve as control accounts, shown alongside their Twitter handle, their corresponding influencer, number of followers at the time of corresponding influencer's deplatforming, and number of tweets referring to them that we collected.

each α. For Alex Jones, we also included hashtags that referenced *Infowars* because Alex Jones owns and runs Infowars, and many posts linked them together.

Note that the coders familiarized themselves with the ideas and conspiracy theories promoted by each deplatformed user, α, by reading media articles and a sample of Twitter posts about him. Following this, for each hashtag, the coders inspected a random sample of 10 tweets in α-D that included that hashtag. Hashtags that did not explicitly reference α were removed. For example, for Alex Jones, we removed the frequently co-occurring hashtags "#MAGA," "#Obama," and "#China." Finally, the two coders came to a consensus on their differences through mutual discussions.

Next, we collected tweets containing those new hashtags, adding them to α-D. This sampling and data collection procedure were repeated until we could no longer find any more hashtags referencing α. Note that we included both supporting and opposing hashtags in this data collection stage, which let us examine changes in the overall volume of discussions around influencers. However, for answering RQ 3, as we will discuss in section 3.4, we limited our data collection to only the tweets posted by influencers' supporters.

Table 2 shows the complete list of keywords we used to collect tweets for each deplatformed influencer. The column of *# Tweets* in Table 1 lists the number of tweets corresponding to each influencer that we collected. For each α, we further divided α-D into tweets posted before and after the deplatforming of α.

Further, to ensure that the trends we observe are not the effects of Twitter-wide patterns, we collected data for a set of control accounts that did not get deplatformed. Specifically, we collected tweets that referred to Drudge Report, Steven Crowder, and Gateway Pundit, which served as

controls for Alex Jones, Milo Yiannopoulos, and Owen Benjamin respectively. Our selection of controls was based on the similarities of these accounts in terms of their rhetoric. Finding an exact influencer match to serve as controls for the deplatformed influencers was not possible because each influencer we study was unique in their views and personality. Therefore, we present this as a best effort control analysis that simply serves to check whether the temporal trends in our results are observed across all influencers, regardless of whether they got deplatformed.

To identify potential control candidates, we collected user handles that most frequently appeared in tweets for a given deplatformed user. We filtered out handles for accounts that did not have within +/-50% of the followers of the deplatformed user at the time of their deplatforming. We further removed accounts directly associated with the deplatformed user to avoid potential confounds (e.g. accounts of Infowars employees). We then qualitatively analyzed the content of the potential control accounts and selected for those most similar in their rhetoric to the deplatformed user within our limited pool of control candidates. Table 3 shows the number of tweets we collected for each control account. It also indicates their follower counts around the time of corresponding influencer's deplatforming.[8]

3.3 Data for RQ2: Tweets Referencing Ideas Promoted by Deplatformed Influencers

To analyze changes in the spread of offensive ideas and conspiracy theories associated with the deplatformed influencers in our sample, for each influencer α, we collected all the n-grams (n = 1, 2, 3) appearing in α-D and created a vocabulary of these n-grams. Next, we compared their frequencies in α-D with their frequencies in other influencers' corpora. Specifically, for each n-gram in our vocabulary, we calculated the log odds ratio of the probability of that n-gram occurring in a tweet in α-D and the probability of it occurring in the combined dataset of other influencers. We then sorted the n-grams in decreasing order of their log odds ratios and manually reviewed the top 300 n-grams to identify keywords that represented the offensive ideas and/or conspiracy theories that were associated with α. We used our knowledge of the influencers to guide this selection. For example, although the top n-grams for Alex Jones included *#trump*, *#usa*, and *#nytimes*, we did not select them for our analysis because they were clearly not offensive and/or not relevant to Alex Jones alone. Through this process, we selected the following n-grams for our analyses: for Alex Jones: *bilderberg*, *#chemtrails*, *deepstate*, *#followthewhiterabbit*, *#qanon8chan*, *#pedogate*, *#pedowood*, *#pizzagate*, and *sandy hook*; for Milo Yiannopoulos: *#stopislam*, *#fuckislam*, *#cuck*, *#islamistheproblem*, *regressives*, *faggots*, *triggering*, and *antisjw*; and for Owen Benjamin: *#soyboy*, *the n word*, and *blackballed*.

To analyze how the use of these keywords evolved on Twitter, we collected all the tweets containing each keyword that were posted in the period 6 months before to 6 months after the corresponding influencer's deplatforming.

3.4 Data for RQ3: All Tweets Posted by Supporters of Deplatformed Influencers

To identify each influencer's supporters and collect their tweets, we began by identifying, for each influencer α, Twitter users in the α-D dataset who supported α. We collected users who had at least 10 posts in α-D before α got deplatformed. This filtering helped us focus on users who had sufficient awareness of α and his views. We also filtered α-D to contain only tweets posted by these users in this stage.

Next, we conducted a descriptive analyses of tweets in α-D posted by these users. Specifically, we identified the most frequently occurring n-grams (n = 1, 2, 3) after excluding stop-words. Our manual review of these n-grams indicated that for Milo Yionnapoulus and Owen Benjamin, most of the frequently occurring n-grams were supportive (e.g., "#freemilonow", "#milosexual"); for Alex

[8]We obtained these counts using snapshots from the Internet archive site, WaybackMachine.

Jones, frequently occurring *n*-grams included both supporting and opposing terms. Driven by this early insight, we randomly sampled 100 users each from Milo Yiannopoulus and Owen Benjamin datasets. Two researchers manually coded each user as 'supporter' or 'opponent' by reviewing the tweets of that user in the corresponding α-D. After disagreements were resolved, this coding labeled 93 of 100 users from the Milo Yiannapoulus sample and 98 of 100 users from the Owen Benjamin sample as supporters.

Since this coding identified such a high proportion of users as supporters, we collected the tweets of *all* users in the Milo Yiannapoulus and Owen Benjamin datasets for our analyses of supporters' behavior. Note that we included only those users who posted at least 10 times pre-deplatforming during this data collection, which spanned the period six months before to six months after the corresponding influencer's deplatforming event. We acknowledge that some of these users may not have been supporters but, as our preceding analyses indicate, such users comprised only a small proportion of the population.

Our *n*-gram analysis indicated that the Alex Jones dataset contained tweets posted by both supporters and opponents of Alex Jones. To better identify supporters, we first identified a small set of supporters and opponents in the Alex Jones dataset and then trained a machine learning classifier to label more supporters. Appendix A describes the details of this process. In total, through this process, we labeled 2,935 users as Alex Jones supporters. Next, we collected *all the tweets* posted by these users in the period six months before to six months after the deplatforming of Alex Jones.

For each influencer, the columns *# Supporters* and *# Supporters Tweets* in Table 1 show the number of their supporters we identified and the number of tweets these supporters posted.

It is possible that our data collection coincided with organized information operations that occurred on Twitter, either using state-operated campaigns [4, 42] or algorithmic bots [27, 33]. Following [40], we compared our data with a list of 83,669 state controlled accounts published by Twitter between Oct 2018 and Feb 2021 [93]: none of these accounts appeared in our datasets described in the previous subsections. Determining whether the accounts in our dataset are bot accounts or not is a difficult, complex and evolving problem [31, 52] and is beyond the scope of this research[9]. We note that much of our analyses would remain relevant even if the accounts in our dataset are part of organized campaigns or bots because the accounts' posts were accessible to and had an impact on observers.

4 METHODS

Figure 1 shows the methodological framework we use to answer our research questions. In brief, we analyzed how conversations about influencers changed after they were deplatformed, explored how deplatforming changed the spread of ideas associated with influencers, and examined how overall posting activity and toxicity levels of influencers' supporters changed after deplatforming.

We divided all our datasets — tweets referencing influencers, tweets referencing influencers' offensive ideas, and tweets posted by influencer's supporters—into time windows of 1 day each. We assigned an index w_i to each time window, with w_0 denoting the time window beginning at the date of the corresponding influencer's deplatforming (i.e., 09/06/2018 for Alex Jones, 07/19/2016 for Milo Yiannopoulos, and 04/05/2018 for Owen Benjamin).

4.1 Operationalizing Different Indicators of User Behavior

We now describe how we operationalized user behavior to answer our three research questions.

[9]We used a popular bot detection service, Botometer, but our manual analysis of a sample of results from this service revealed many errors. We therefore chose not to rely on this service to make claims about the activity levels of bots in our datasets.

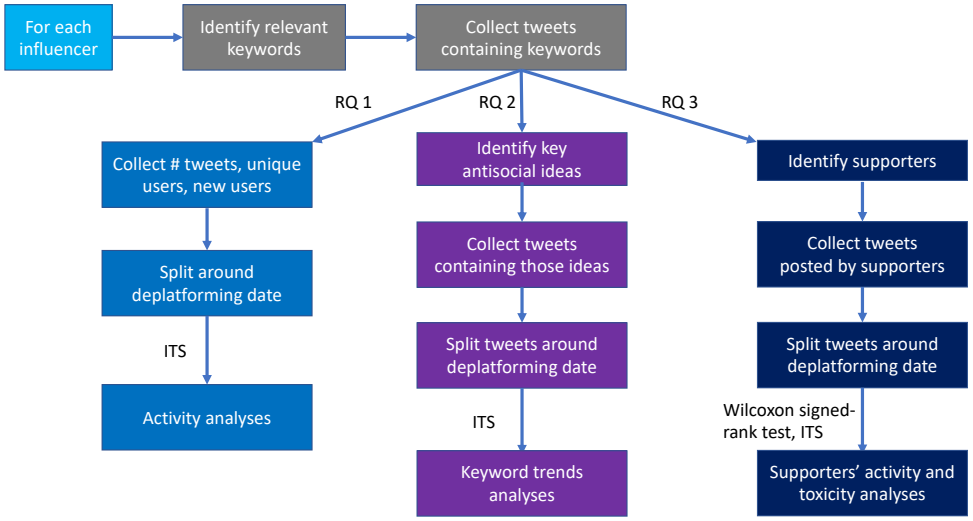


Fig. 1. Methodological framework used in this study to examine how deplatforming affects (1) change in activity around influencers (RQ 1), (2) change in the spread of ideas associated with influencers (RQ 2), and (3) change in overall activity and toxicity levels of their supporters (RQ 3).

4.1.1 RQ1: Change in Conversations about Deplatformed Influencers. We measured three different aspects of conversations about influencers:

- *Posting activity levels.* We measured the posting activity levels for each influencer α using the volume of tweets in the corresponding dataset α-D posted in each time window w. For this, we iterated through each tweet in α-D and assigned it to the corresponding time window based on the timestamp indicating when the tweet was made. Next, we aggregated all tweets that were assigned to each time window. The total number of tweets contained in each time window represented the activity levels corresponding to α.
- *Number of unique users.* We measured the number of unique users posting about each influencer α in each time window w by examining the tweets assigned to that window. We created a set of distinct users who posted these tweets and calculated the set length.
- *Number of new users.* For each time window w in the dataset for each influencer α, we defined *new users* as users who had never posted about α before w. To distinguish new users from preexisting ones, we set aside an initial buffer period of seven days for which we did not compute new users. We collected all unique users who posted in the first seven days in a set *seenUsers*. Starting from the eighth day, we collected unique users posting about α in each time window, *currentUsers(w)*, and curated new users in that time window, *newUsers(w)*, by compiling those users who were not in *seenUsers* (i.e., *newUsers(w)* is the set difference, *currentUsers(w)* − *seenUsers*). We iterated through each subsequent time window, adding new users seen in that window to the set *seenUsers* (i.e., *seenUsers* = *seenUsers* ∪ *newUsers(w)*). We repeated this process for each deplatformed influencer in our sample.

4.1.2 RQ2: Change in Spread of Offensive Ideas Associated with Influencers. As discussed in Section 3.3, we identified a set of offensive ideas, operationalized through keywords, associated with each

influencer and collected tweets referencing those keywords. We measured two key aspects of the spread of each offensive idea: (1) number of tweets mentioning that idea, and (2) number of unique users mentioning that idea. We again divided our tweets mentioning each idea into time windows of one day each. Similar to our process in the previous section, we measured volume of tweets and number of unique users referencing each idea in each time window.

4.1.3 RQ3: Change in Overall Behavior of Supporters. As in the previous section, we divided each supporter's tweets into time windows of one day each. We assigned an index w_i to each time window, with w_0 denoting the time window beginning at the date of deplatforming of the corresponding influencer. We measured two important aspects of supporters' behavior:

- *Posting activity levels.* We measured the posting activity levels of each supporter, s, using the number of tweets s posted in each time window w. For this, we iterated through each tweet of s, and assigned it to the corresponding time window based on its timestamp. Next, we aggregated all tweets assigned to each time window. The total number of tweets contained in each time window represented the activity level of s in that window.
- *Toxicity levels.* The influencers we study are known for disseminating offensive content. Can deplatforming this handful of influencers affect the spread of offensive posts widely shared by their thousands of followers on the platform? To evaluate this, we assigned a toxicity score to each tweet posted by supporters using Google's Perspective API. This API leverages crowdsourced annotations of text to train machine learning models that predict the degree to which a comment is rude, disrespectful, or unreasonable and is likely to make people leave a discussion.[10] Therefore, using this API let us computationally examine whether deplatforming affected the quality of content posted by influencers' supporters. Through this API, we assigned a *Toxicity* score and a *Severe Toxicity* score to each tweet. The difference between the two scores is that the latter is much less sensitive to milder forms of toxicity, such as comments that include positive uses of curse words.[11] These scores are assigned on a scale of 0 to 1, with 1 indicating a high likelihood of containing toxicity and 0 indicating unlikely to be toxic. For analyzing individual-level toxicity trends, we aggregated the toxicity scores of tweets posted by each supporter s in each time window w.

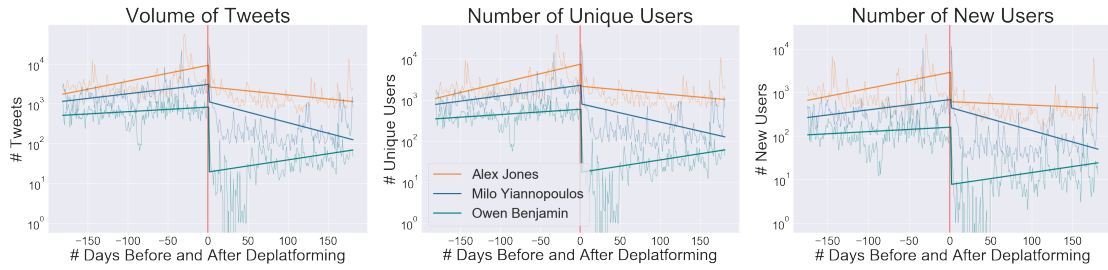
We acknowledge that detecting the toxicity of text content is an open research problem and difficult even for humans since there are no clear definitions of what constitutes inappropriate speech [103]. Therefore, we present our findings as a best-effort approach to analyze questions about temporal changes in inappropriate speech post-deplatforming.

4.2 Causal Inference

The causal inference question we explore is whether the deplatforming of offensive influencers like Alex Jones leads to a decrease in radicalization on Twitter as measured through the lens of posting activity levels, toxicity levels, etc. To determine whether the post-deplatforming effects observed could instead be explained by random chance, we use a causal inference strategy called Interrupted Time Series (ITS) Regression Analysis [9]. ITS lets us measure the causal effects of deplatforming, accounting for ongoing temporal trends for different indicators of user behavior.

In an ITS analysis, we track the dependent variable (e.g., posting activity level) over time and use regression to determine if a *treatment* at a specific time (*treatment* for this study is the deplatforming event) caused a behavior change (e.g., change in posting activity levels). The ITS regression models the behavior of the pre-treatment time series to predict how the series would have looked if the

[10] https://github.com/conversationai/perspectiveapi/blob/master/2-api/models.md
[11] https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages



(a) Posting Activity Levels (b) Number of Unique Users (c) Number of New Users

Fig. 2. Variations in (a) posting activity levels, (b) number of unique users, and (c) number of new users posting about Alex Jones, Milo Yiannopoulos, and Owen Benjamin before and after their deplatforming. Results show that for each metric and each influencer, activity levels declined after deplatforming.

treatment had not been applied. This is usually done by regressing the behavior data on time and including in the regression a binary indicator variable for the post-treatment time periods (e.g., [16, 17]). It is important to regress the behavior on time; otherwise, we could easily misinterpret a steadily increasing (or decreasing) time series as a treatment effect when comparing the averages of behavior prior to and post treatment [9]. In this sense, an ITS regression allows us to claim that the behavior actually changed at the treatment time (time of deplatforming) instead of simply reflecting a general trend.

5 RESULTS

5.1 RQ1: Change in Discussions about Deplatformed Influencers

For analyzing the change in discussions, we employed a separate poisson regression model for analyzing data corresponding to each metric (volume of posts, number of unique posters, new user influx) and each of the influencers in our study:

$$y_t = c + \alpha t + \beta i_t \tag{1}$$

where t is the date, which takes values between -180 and $+180$ and equals 0 on the day of deplatforming; y_t is the statistic we are modeling; and i_t is an indicator variable equal to 1 for days following deplatforming (i.e., $t > 0$), and 0 otherwise.

5.1.1 Temporal Changes in the Volume of Posts Following Deplatforming. Figure 2(a) shows the temporal trends in posting activity levels corresponding to the three deplatformed accounts from our sample. As this figure shows, we observe dramatic drops after deplatforming in each case. To causally attribute these drops to deplatforming events, we performed ITS analyses of posting volumes for each deplatformed influencer. Table 4 shows the results of these analyses, which show that deplatforming significantly reduced the number of posts referencing each influencer—by 91.77% on average.

5.1.2 Temporal Changes in the Number of Unique Posters Following Deplatforming. Next, we analyze whether the number of unique users posting about the three influencers in our sample changed after deplatforming. Figure 2(b) shows the temporal trends in the number of users referencing each influencer. We found that in each case, the number of users precipitously declined after deplatforming. To verify that these drops are attributable to deplatforming and not to existing temporal trends, we conducted ITS analyses on the number of unique users. Table 4 presents the results of these analyses, which suggest that even after controlling for linear temporal trends,

Table 4. Interrupted Time Series (ITS) Regression results for volume of tweets, number of unique users, and number of new users corresponding to each deplatformed influencer in our sample. We include the β coefficients for the post-treatment indicator from ITS regression and the percentage change in each metric caused by deplatforming. In each case, β is significant with p<.001. Results show a highly significant causal effect of deplatforming on the substantial decrease in activity levels.

	# Tweets		# Unique Users		# New Users	
Influencer	**Coeff**	**% Change**	**Coeff**	**% Change**	**Coeff**	**% Change**
Alex Jones	-2.06	-87.25%	-1.88	-84.74%	-1.87	-84.59%
Milo Yiannopoulos	-2.35	-90.46%	-2.1	-87.75%	-2.12	-88.00%
Owen Benjamin	-3.74	-97.62%	-3.23	-96.04%	-3.32	-96.38%

deplatforming was associated with a reduction in the number of unique posters tweeting about influencers: across the three influencers, deplatforming reduced the number of unique posters by an average of 89.51%.

5.1.3 Temporal Changes in New User Influx Following Deplatforming. Influential public figures who spread conspiracy theories and post offensive content can attract an increasing number of *new* people to such content. Therefore, it is vital to consider whether deplatforming such figures can reduce the influx of new users who are attracted to them. Figure 2(c) shows the temporal trends in the number of new users referencing the three influencers we study. This figure shows that for each deplatformed influencer, the number of new users posting about them dropped steeply post-deplatforming.

Again, to verify that these drops can be attributed to deplatforming and not just to existing temporal trends, we conducted ITS analyses on the number of new users. Table 4 shows the results of these analyses, which reveal that in all cases, deplatforming reduced the number of new users tweeting about the corresponding influencer. Across the three influencers, deplatforming reduced the number of new users by an average of 89.65%. This suggests that deplatforming can help reduce the influx of new users who are attracted to offensive influencers.

5.1.4 Further Analysis of the Sudden Drops. Noticing the sudden drops in all the measured metrics for the three deplatformed accounts, we further analyzed the mechanisms that could help explain these drops. First, looking at the raw counts of these metrics around the dates of deplatforming, we noticed that there was a sudden increase in posting activity right before the deplatforming events that dissipated quickly after the bans. Analyzing a sample of these posts, we found that they referred to the unpopular offline events that these influencers were involved in. Such events also could have ultimately contributed to the deplatforming decisions. For example, Alex Jones tweeted out a video of himself heckling the CNN reporter Oliver Darcy, and Twitter's announcement of banning Jones noted:

> "*Today, we permanently suspended @realalexjones and @infowars from Twitter and Periscope. We took this action based on new reports of Tweets and videos posted yesterday that violate our abusive behavior policy, in addition to the accounts' past violations.*"[12]

In light of such offline events, many opponents of these influencers called for banning them from Twitter. Such calls promptly reduced after deplatforming occurred, which partially explains the sudden drop in posting activity metrics. We also found that once banned, the number of tweets mentioning these influencers' handle drastically reduced (see Table 5). We estimate that this happened because once an influencer is banned, his tweets and user handle are no longer accessible

[12]https://twitter.com/TwitterSafety/status/1037804427992686593

	Before	After
Alex Jones	519.6	3.2
Milo Yiannopoulos	92.9	41.7
Owen Benjamin	115.6	1.9

Table 5. Mean number of tweets daily mentioning each influencer before and after their deplatforming

	# Tweets		# Unique Users		# New Users	
Influencer	**Coeff**	**% Change**	**Coeff**	**% Change**	**Coeff**	**% Change**
Alex Jones	-1.01	-63.58	-0.96	-61.71	-1.31	-73.02
Milo Yiannopoulos	-1.66	-80.99	-1.56	-78.99	-1.48	-77.24
Owen Benjamin	-1.55	-78.78	-1.68	-81.36	-1.04	-64.65

Table 6. Comparative Design Interrupted Time Series Regression results for volume of tweets, number of unique users, and number of new users corresponding to each deplatformed influencer in our sample. We include the β coefficients for the interaction between post-treatment indicator and *isDeplatformed* indicator from ITS regression and the percentage change in each metric for deplatformed influencers when compared to control influencers. In each case, β is significant with $p<.001$. Results show a highly significant causal effect of deplatforming on the substantial decrease in activity levels even when compared to similar influencers who were not deplatformed.

on the platform, thereby reducing the ways in which users can directly interact with the influencer. This could also have contributed to the sudden drops in posting metrics corresponding to each influencer.

5.1.5 Accounting for Control Influencers. In order to ascertain that the post-deplatforming changes we observed reflected the effects of deplatforming and not Twitter-wide trends, we conducted an additional set of comparative design interrupted time series analyses by including the posting metrics for control accounts in each of our regression analyses. We added a binary variable *isDeplatformed* to our model that distinguished data entries for deplatformed accounts and control accounts. Table 6 presents the results of these analyses, which suggest that deplatformed influencers saw a significant reduction in their post-deplatforming posting metrics as compared to control influencers.

5.2 RQ2: Change in the Spread of Ideas Associated with Influencers

We next describe our findings on how deplatforming affected the spread of ideas associated with influencers. As discussed in Section 3.3, we manually curated the list of ideas to test for this analysis. This is, of course, not a comprehensive list of ideas popularized by influencers; however this analysis offers some preliminary insights into how deplatforming impacts not just the popularity of influencers (as seen above) but also of the offensive topics they promote.

As described in Section 3.3, we collected for each keyword all tweets containing that keyword posted in the period six months before to six months after the deplatforming of the corresponding influencer. Similar to our analyses in the previous section, for each keyword, we conducted a separate ITS analysis on posts referencing that keyword to measure how deplatforming affected its use. We use the same regression model as in equation 1 to model the metrics corresponding to each keyword. Table 7 shows the results of these regression analyses for daily volume of tweets and number of unique users. It also displays the mean daily pre-deplatforming and post-deplatforming values for the two metrics. Figure 3 shows these trends for a sample of the keywords.

We found that deplatforming reduced the tweet volume and the number of users referencing a majority of the analyzed keywords; this demonstrates the effectiveness of deplatforming in

Table 7. Results analyzing the spread of ideas associated with Alex Jones, Milo Yiannopoulos and Owen Benjamin. We show the pre-deplatforming and post-deplatforming mean (μ) daily (1) volume of tweets mentioning these keywords, and (2) number of unique users posting these tweets. The β coefficient and p-value for the post-treatment indicator from ITS regression, as well as the percentage change in tweet volume and number of unique users caused by deplatforming are also included. These results suggest that the spread of most of these ideas was reduced due to deplatforming. Here, p<.05: *, p<.01: **, p<.001: ***

Influencer	Keyword	Tweets Volume				Unique Users			
		Pre μ	Post μ	Coeff	% Change	Pre μ	Post μ	Coeff	% Change
Alex Jones	**bilderberg**	467	585	0.06***	6.18%	208	168	-0.51***	-39.95%
	#chemtrails	126	116	0.06**	6.21%	80	78	0.1***	10.52***%
	deepstate	3882	3064	-0.22***	-19.75%	3099	2470	-0.24***	-21.34%
	#followthewhiterabbit	125	41	0.55***	73.33%	46	16	0.2***	22.14%
	#qanon8chan	159	58	0.64***	89.64%	63	20	0.38***	46.23%
	#pedogate	292	98	-1.03***	-64.3%	175	65	-1.16***	-68.65%
	#pedowood	61	27	-2.17***	-88.58%	38	18	-2.27***	-89.67%
	#pizzagate	270	117	-0.66***	-48.31%	184	90	-0.74***	-52.29%
	sandy hook	727	517	-0.18***	-16.47%	616	428	-0.23***	-20.55%
Milo Yiannopoulos	**#stopislam**	466	167	-0.03*	-2.96%	247	52	-0.47***	-37.5%
	#fuckislam	16	10	-1.2***	-69.88%	12	8	-0.92***	-60.15%
	#cuck	28	44	0.11**	11.63%	19	31	0.12**	12.75%
	#islamistheproblem	115	42	-0.45***	-36.24%	61	30	-0.37***	-30.93%
	regressives	84	88	-0.07**	-6.76%	39	42	-0.08**	-7.69%
	faggots	1865	1494	-0.04***	-3.92%	1619	1281	-0.04***	-3.92%
	triggering	750	769	-0.88***	-58.52%	702	721	-0.88***	-58.52%
	antisjw	8	14	0.24***	27.12%	8	12	0.23**	25.86%
Owen Benjamin	**#soyboy**	28	21	-0.59***	-44.57%	24	19	-0.62***	-46.21%
	the n word	824	897	0.17***	18.53%	688	769	0.16***	17.35%
	blackballed	117	94	-0.02	-1.98%	110	89	-0.02	-1.98%

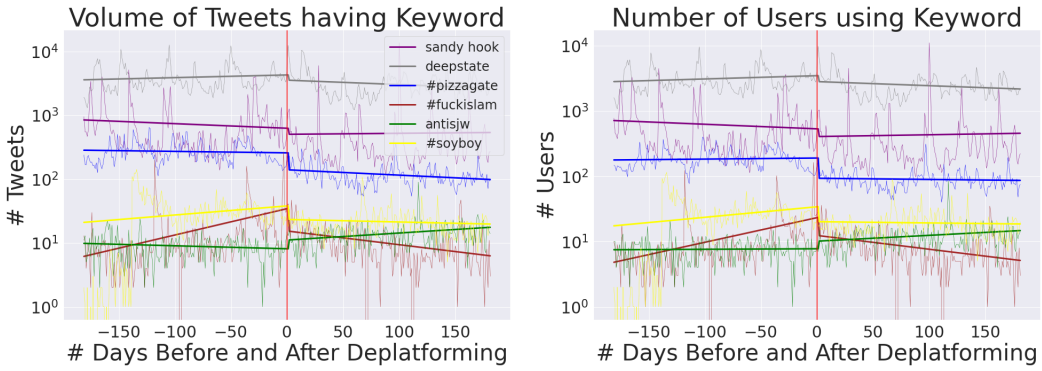
reducing the spread of toxic content as well as the number of individuals spreading offensive ideas. However, we also found a number of exceptions, including *#chemtrails*, *#qanon8chan* and *the n word*, that increased in their frequency after deplatforming. This suggests that in the aftermath of deplatforming, some offensive ideas and conspiracy theories may still continue to gain traction by the supporters of deplatformed influencers. Another explanation could be that although these keywords were used more frequently by supporters of their corresponding deplatformed influencer than by supporters of other influencers in our sample (which is why we selected them for our analyses (Section 3.3)), these keywords may also be popular in other contexts, and an increase in their use may not be related to deplatforming.

Overall, these results suggest that deplatforming helped reduce the spread of a majority of offensive ideas associated with deplatformed users. This highlights how deplatforming may help improve the quality of conversations on the platform.

5.3 RQ3: Change in Supporters' Behavior

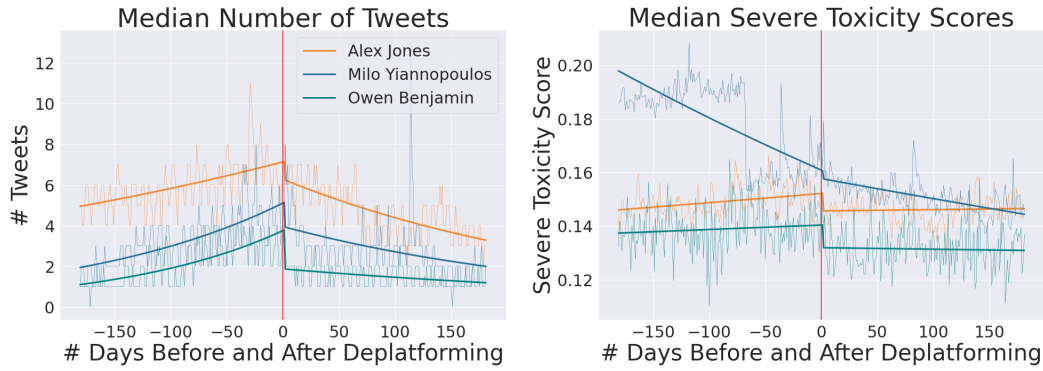
We next analyze changes in the overall activity of supporters.

5.3.1 Change in Posting Activity Levels of Supporters Post-Deplatforming. Figure 4(a) shows the temporal trends in median posting activity levels for the supporters of the three influencers we studied. Based on this figure, we observe drastic drops in posting levels after deplatforming in each case. We further confirmed this finding through Wilcox signed-rank tests. For each influencer, we created a list of pre-deplatforming posting levels of their supporters and another list containing post-deplatforming posting levels of each supporter. Next, we conducted a Wilcoxon signed-rank test between these two lists to determine the effects of deplatforming on posting levels. As shown in Table 8, these analyses highlight a statistically significant median decrease in posting activity levels



(a) Posting Activity Levels (b) Number of Unique Users

Fig. 3. Variations in daily (a) posting activity levels and (b) number of unique users using keywords selected at random from the list of ideas we analyze. Results show a decrease in activity levels after deplatforming for most keywords.



(a) Median Posting Activity Levels (b) Median Severe Toxicity Levels

Fig. 4. (a) Median posting activity levels, and (b) median *Severe Toxicity* scores of the supporters of Alex Jones, Milo Yiannopoulos and Owen Benjamin before and after their deplatforming. Results show a decrease in both activity and toxicity levels after deplatforming each influencer.

of each influencer's supporters after deplatforming.[13] Across the three influencers, we observed a median decline of 12.59% in the volume of tweets posted by their supporters on average. This suggests that deplatforming an influencer may drive their most ardent supporters away from the platform.

For each influencer, we also conducted an ITS regression analysis for each supporter and calculated the number of supporters who significantly[14] changed their posting levels after deplatforming. Table 9 shows the results of these analyses. These results indicate that for each influencer, the number of supporters who significantly increased their posting activity due to deplatforming is

[13]To test whether deplatforming led to short-term disruptions that affect these results, we also conducted all Wilcoxon signed-rank tests in our paper on the same datasets but excluded the tweets posted in the period one month before to one month after deplatforming. These analyses produced qualitatively similar results.
[14]Significance measured at p<.01.

Table 8. Results for Wilcoxon signed rank tests comparing the pre- and post-deplatforming posting activity levels of supporters of Alex Jones, Milo Yiannopoulos and Owen Benjamin. We include the median number of pre- and post-deplatforming tweets, the percentage median change in number of tweets, as well as the W, *p* and *z*-values. Results show that overall activity levels of supporters declined significantly in each case after deplatforming. Here, p<.001: ***

	Pre Median	Post Median	Median diff %	W	z
Alex Jones	1622	1322	-9.31%	1558191***	-12.96
Milo Yiannopoulos	1235	1033	-9.96%	6471036***	-15.65
Owen Benjamin	819	573	-18.50%	14701***	-5.46

Table 9. Summary of ITS regression results for posting activity levels of supporters of Alex Jones, Milo Yiannopoulus and Owen Benjamin. For each influencer, we show the total number of supporters, the total number of supporters who significantly increased and decreased their posting levels after deplatforming, and the median percentage change in posting levels caused by deplatforming. For each influencer, fewer supporters significantly increased their posting levels after deplatforming than decreased them.

	# Supporters	Sig. Increase (Median % Change)	Sig. Decrease (Median % Change)
Alex Jones	2,935	934 (73.33%)	1,234 (-51.32%)
Milo Yiannopoulos	5,827	1,845 (91.55%)	2,506 (-59.55%)
Owen Benjamin	304	62 (108.55%)	168 (-67.69%)

lower than the number of supporters who significantly decreased their posting activity. However, the former group showed a higher median increase in their posting activity levels than the decrease in posting activity levels shown by the latter group in each case. This shows that deplatforming did not deter all supporters, and, indeed, fired up a small set of supporters to post more than before. Still, as our Wilcoxon signed-rank tests show, the overall effect was a decline in posting levels of supporters after deplatforming.

5.3.2 Change in Toxicity Levels of Supporters Post-Deplatforming. We next analyzed the change in toxicity levels of each influencer's supporters post-deplatforming. As described in section 4.1.3, we measured toxicity levels using two Perspective API scores: *Toxicity* and *Severe Toxicity*. We obtained qualitatively similar results for both *Toxicity* and *Severe Toxicity* scores, so we present here only the results of *Severe Toxicity* scores for brevity.

Figure 4(b) shows the temporal trends in median *Severe Toxicity* scores of the supporters of the three influencers. Each point in these plots represents for a date the median of the average toxicity scores of all supporters on that date. This figure indicates that for all three influencers, supporters' toxicity levels declined after deplatforming.

Next, we statistically measured whether there were significant differences in the toxicity scores of supporters' posts after the influencers were deplatformed. For each influencer, we created a list of pre-deplatforming average *Severe Toxicity* scores for each of their supporters and another list containing post-deplatforming average *Severe Toxicity* scores for each supporter. Next, we conducted Wilcoxon signed-rank tests between these two lists to evaluate the effects of deplatforming on *Severe Toxicity* scores. Table 10 shows the results of these tests; they reveal that toxicity levels of each influencer's supporters significantly dropped after deplatforming: across the three influencers, we observed a median decline of 5.84% in the toxicity levels of their supporters on average. This suggests that absent the influence of deplatformed accounts, supporters may reduce their toxicity and improve their behavior.

Table 10. Results for Wilcoxon signed rank tests comparing the pre- and post-deplatforming *Severe Toxicity* scores of supporters of Alex Jones, Milo Yiannopoulos and Owen Benjamin. We include the median pre-deplatforming and post-deplatforming *Severe Toxicity* scores, the percentage median change in these scores, as well as the W, *p* and *z*-values. These results show that supporters' toxicity levels significantly reduced after deplatforming in each case. Here, p<.001: ***

	Pre Median	Post Median	Median diff %	W	*z*
Alex Jones	0.162	0.159	-1.23%	1805247***	-5.71
Milo Yiannopoulos	0.192	0.168	-12.50%	2039385***	-48.31
Owen Benjamin	0.158	0.151	-3.80%	14741***	-4.76

Table 11. Summary of Interrupted Time Series regression results for Severe Toxicity scores of supporters of Alex Jones, Milo Yiannopoulus and Owen Benjamin. For each influencer, we show the total number of supporters studied, the total number of supporters who significantly increased and decreased their toxicity levels after deplatforming, and the median percentage change in Severe Toxicity scores caused by deplatforming. For each influencer, most supporters did not significantly change their toxicity scores. In each case, fewer supporters significantly increased their toxicity levels than decreased them.

	# Supporters	Sig. Increase (Median % Change)	Sig. Decrease (Median % Change)
Alex Jones	2,935	87 (5.13%)	172 (-3.92%)
Milo Yiannopoulos	5,827	341 (6.18%)	506 (-4.88%)
Owen Benjamin	304	6 (7.79%)	8 (-6.29%)

We also analyzed individual-level toxicity trends. For each influencer, we conducted an ITS analysis for each supporter and calculated the number of supporters who significantly[15] changed their toxicity levels after deplatforming. Table 11 shows the results of these analyses, which indicate that for each influencer, most supporters did not significantly change their toxicity scores. Still, the number of supporters who significantly increased in toxicity after deplatforming is lower than the number of supporters who significantly decreased in their toxicity. We also found that the former group showed a slightly higher median increase in their toxicity levels than the decrease in toxicity levels shown by the latter group in each case. This suggests that deplatforming worsened the posting outcomes of a small group of supporters. Still, as our prior Wilcoxon signed rank test results show, although deplatforming did not reform all supporters, its overall effect was to reduce toxicity after deplatforming.

6 DISCUSSION

Taken together, our findings show a robust effect of deplatforming offensive influencers on disrupting the growth of communities formed around them. We next discuss the theoretical and design implications of this work, note its limitations, and suggest directions for future work.

6.1 Deplatforming is an Effective Strategy to Reduce Offensive Influencers' Impact and Lessen Toxic Rhetoric

We note that deplatforming is an organizational decision and as researchers unaffiliated with Twitter, we do not have any insights into how the platform decides to deplatform some accounts but not others. Instead, our focus in this paper is on examining the long-term effects of deplatforming on posting behaviors on Twitter.

[15]Significance measured at p<.01.

Conversations around influencers are reduced. Before this work, it was not clear whether or how deplatforming would influence the levels of posting activity about influencers. Certainly, these influencers were still actively promoting their views elsewhere. In fact, their deplatforming on Twitter was widely covered by multiple news outlets and provided them greater visibility, which could lead to a Streisand effect of drawing more attention to censored individuals. However, our results show that deplatforming significantly reduced the number of postings about these influencers. Additionally, the number of new users and unique users posting about them declined dramatically. Thus, we conclude that deplatforming helped reduce the overall impact of these influencers on the platform.

Spread of offensive ideas associated with influencers are reduced. We analyzed the spread of many offensive ideas associated with the deplatformed influencers. Our ITS analyses show that even after controlling for temporal trends, deplatforming helped reduce the spread of many of these anti-social ideas and conspiracy theories. This suggests that deplatforming diminishes not just the influence of banned individuals, but also of their ideas.

Activity and toxicity levels of supporters are reduced. Deplatforming influencers could have fired up their supporters and raised their posting activity and toxicity levels. Extremist communities are often motivated by alleged aggrieved victimhood to draw greater support [85], and the deplatforming of their leaders could have been a call to arms. However, our analysis of the long-term activity of their supporters reveals that deplatforming helped reduce their overall posting activity and toxicity levels. Since the influencers we studied were deplatformed at different times, it is unlikely that the changes we observed are reflective of Twitter-wide trends. Thus, deplatforming can have wide-ranging and longer-term positive consequences on the platform health.

6.2 Platforms Must Deplatform Influencers Who Promote Offensive Speech Even in the Face of Lost Advertising Dollars

We found that deplatforming influencers reduced the posting activity levels of hundreds of their supporters. Therefore, it might not be in the financial interests of platforms to conduct deplatforming. Many critics have raised concerns about the financial benefits from advertising dollars that are potentially tied to allowing toxic content to remain on these platforms [15, 28]. Indeed, some evidence suggests that platforms appears willing to bend their rules for popular extremist influencers. According to leaked internal Facebook materials reviewed by NBC, Facebook executives removed "strikes" from the accounts of several high-profile influencers, like *Charlie Kirk* and *Diamond and Silk*, who had shared viral misinformation [86]. Similarly, a Bloomberg report suggested that YouTube was hesitant to police videos shared on its platform because it feared that doing so would bring supporter engagement down [7]. However, platforms should recognize that when they allow people who promote toxic speech to spread their views in the name of free speech, they are degrading women, minorities and other vulnerable groups and minimizing their dignity. Further, and possibly of more concern to these platforms, these groups may abandon social media use if it continues to represent toxic rhetoric. In light of these concerns, it is vital that platforms clarify their commitment to respecting the dignity of all their users and deplatform offensive influencers when it is appropriate.

6.3 Platforms must defend against the second-order harms of deplatforming

We found that deplatforming increased the prevalence of some offensive ideas we tested in this study. Therefore, platforms should be cautious that in the aftermath of deplatforming, certain user groups may increase their spread of vitriolic ideas and conspiracy theories associated with the banned influencer. Careful moderation of posts containing these ideas may help reduce the

negative consequences of deplatforming. For this, our methods for identifying relevant keywords and measuring their temporal trends may provide a useful guideline.

In a similar vein, we found that although deplatforming helped reduce the overall activity and toxicity levels of supporters, a small group of supporters significantly increased both their activity and toxicity levels. Thus, platforms should attend to how deplatforming may impact the activities of other users associated with the banned accounts and regulate their activities when necessary. Our data collection and methodological approach may prove helpful in identifying supporters and analyzing their change in behaviors. We also observed that far fewer supporters changed their toxicity levels after deplatforming than those who changed their posting activity levels (Tables 9 and 11). This suggests that in the face of platform interventions, users may not reform their behavior as much as they change their posting levels.

6.4 Toward Building a Theory for Effective Moderation Interventions

Given the effectiveness of deplatforming that we highlight, we hope that this research serves as a lever to open a discussion about determining appropriate thresholds for de-platforming. The proposed methods and data cleaning procedures are generic and can be generalized to evaluate deplatforming in other contexts, e.g., on Facebook and Instagram. Our analyses employ new metrics that may be used to evaluate the second-order effects of moderation interventions. In conjunction with prior literature [5, 17, 47, 49, 54, 84] that evaluates the effectiveness of moderation strategies like banning, quarantining, and offering explanations for post removals, this research contributes to building a theory [54, 89, 90] that prescribes for community managers which moderation interventions they should deploy and under what circumstances.

6.5 Limitations and Future Work

Focus on effects within Twitter. We looked at the influence of deplatforming controversial influencers only on the platforms where they were banned. It is likely that on being banned, these influencers migrate to other platforms and continue to propagate their ideas. Indeed, prior research notes that after their deplatforming, Alex Jones and Milo Yiannopoulos moved to Gab [77]. Additionally, Alex Jones also asked his supporters to migrate to Infowars, a fake news website he owns. In future work, it would be fruitful to examine the effects of deplatforming on such migrations.

For the cases we studied, we observed that multiple mainstream platforms like Facebook, Twitter and YouTube deplatformed these influencers in quick succession. Although we have conducted our analyses only on Twitter, it seems likely that migration of influencers' supporters to other online spaces would be severely hampered when multiple popular platforms engage in deplatforming. Therefore, we suspect that when platforms learn from one another and deplatform offensive influencers, they can substantially reduce these influencers' ability to propagate toxic speech and recruit supporters. While these influencers may find a home on smaller or more secret platforms, it may still starve them of victims to target online [68]. Indeed, prior research suggests that migration to smaller platforms substantially reduces their audience size and influence [77]. Further research on the factors that influence where deplatformed influencers move, how successfully their supporters move with them on newer platforms, and the extent to which congregating on more obscure online spaces contributes to radicalization would be valuable.

Focus on three influencers. We focus on three offensive influencers as case studies in this work. However, other influencers have been deplatformed. It is likely that the individual factors involved in each case of deplatforming would shape the effects of that deplatforming. Understanding how different factors impact the after-effects of deplatforming is a productive direction for future work. Still, the consistent results from our case studies suggest that deplatforming may generally help reduce the influence of offensive influencers and their thousands of supporters. We focus on only

the supporters of influencers because our data primarily contained users who supported influencers. For other influencers, it might be relevant to also identify the influencers' opponents and examine how their behavior changed after deplatforming.

Other factors that moderate the effects of deplatforming. We note that although deplatforming caused a decline in different posting metrics for each influencer we studied, the levels and trends of decline were quite different for each case study. For example, as Figure 2 shows, posting activity for Milo Yiannopoulos shows a much steeper decrease than the other two influencers. This suggests that there are additional causal factors that moderate the effects of deplatforming. Such factors may include the number of followers an influencer has at the time of deplatforming, how the followers perceive Twitter's sanction of the influencer, whether the influencer has a large following on other platforms, and whether other followers who can easily fill the gap left by the banned influencers already exist on the platform. Analyzing the importance of such moderating factors is a promising direction for future research.

Use of keywords as a proxy for ideas. In section 3.3, we used keywords as a proxy for offensive ideas associated with influencers. Although simple, the use of keywords for this task is effective: we found that users often intentionally employed the hashtags in this list to spread the corresponding ideas. We note that our selection of keywords is manually guided and our analysis is exploratory; with larger datasets (e.g., tweets corresponding to many more influencers), this keyword selection and analysis can be made more robust in future work. In the future, it would be interesting to examine how other linguistic techniques can be employed to capture the use of offensive ideas.

Role of content moderation is unknown. For our analyses, we could collect only those tweets that were not removed by Twitter. It is possible that differences in content moderation before and after deplatforming could explain the differences in activity levels we observed. However, we suspect that this is unlikely since we observed similar patterns for the three influencers we studied, who were deplatformed at different times. Additionally, we could collect data only for those users who were not banned by Twitter at the time of data collection, which occurred many months after deplatforming events. Analyzing new cases of deplatforming may help clarify the extent to which Twitter bans moderate the effects we observed.

7 CONCLUSION

In this paper, we examined the long-term consequences of deplatforming three offensive influencers on Twitter. Our results show that this approach minimized the impact of influencers and their ideas as well as modulated the offensive discourse of their many supporters. We conclude that when used judiciously, deplatforming can be an effective strategy to help detoxify social media. Going forward, additional research is needed to identify the appropriate thresholds for deplatforming and examine its effects on other platforms.

REFERENCES

[1] 2017. Milo Yiannopoulos: Who is the alt-right writer and provocateur? https://www.bbc.com/news/world-us-canada-39026870

[2] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[3] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (may 2017), 211–236. https://doi.org/10.1257/jep.31.2.211

[4] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[5] Sumit Asthana and Aaron Halfaker. 2018. With Few Eyes, All Hoaxes Are Deep. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 21 (Nov. 2018), 18 pages. https://doi.org/10.1145/3274290

[6] Martin Belam. 2018. Milo Yiannopoulos banned from crowdfunding site Patreon. https://www.theguardian.com/technology/2018/dec/06/milo-yiannopoulos-banned-from-crowdfunding-site-patreon

[7] Mark Bergen. 2019. YouTube Executives Ignored Warnings, Let Toxic Videos Run Rampant. https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant

[8] Joe Berkowitz. 2018. Why is Amazon promoting this anti-trans alt right troll's comedy special? https://www.fastcompany.com/90275862/why-is-amazon-promoting-this-anti-trans-alt-right-trolls-comedy-special

[9] 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.

[10] Joseph Bernstein. 2019. Here's How Breitbart And Milo Smuggled Nazi and White Nationalist Ideas Into The Mainstream. https://www.buzzfeednews.com/article/josephbernstein/heres-how-breitbart-and-milo-smuggled-white-nationalism

[11] Nick Bilton. 2019. The Downfall of Alex Jones Shows How the Internet Can Be Saved. https://www.vanityfair.com/news/2019/04/the-downfall-of-alex-jones-shows-how-the-internet-can-be-saved

[12] Lindsay Blackwell, Jill P Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI* 1, CSCW (2017), 24–1.

[13] Porismita Borah. 2014. Does it matter where you read the news story? Interaction of incivility and news frames in the political blogosphere. *Communication Research* 41, 6 (2014), 809–827.

[14] Danah Boyd. 2017. Hacking the attention economy. *Data and Society: Points. Available at: https://points. datasociety. net/hacking-the-attention-economy-9fa1daca7a37* (2017).

[15] Caitlin Ring Carlson, Luc Cousineau, and Caitlin Ring Carlson. 2020. Are You Sure You Want to View This Community? Exploring the Ethics of Reddit's Quarantine Practice. *Journal of Media Ethics* 00, 00 (2020), 1–12. https://doi.org/10.1080/23736992.2020.1819285

[16] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2020. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. arXiv:cs.SI/2009.11483

[17] 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. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 31 (Dec. 2017), 22 pages. https://doi.org/10.1145/3134666

[18] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting Internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3175–3187.

[19] Jonathan Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of Blocked Community Members: Redemption, Recidivism and Departure. In *The World Wide Web Conference*. ACM, 184–195.

[20] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*.

[21] Anna Chung. 2019. How Automated Tools Discriminate Against Black Language. https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db

[22] Danielle Keats Citron. 2014. *Hate crimes in cyberspace*. Harvard University Press.

[23] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing Revenge Porn. *Wake Forest Law Review* 49 (2014).

[24] Nathan Cofnas. 2019. Deplatforming Won't Work. https://quillette.com/2019/07/08/deplatforming-wont-work/

[25] Joseph Cox and Jason Koebler. 2019. https://www.vice.com/en_us/article/nexpbx/facebook-bans-white-nationalism-and-white-separatism

[26] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. (2017).

[27] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*. 273–274.

[28] Elizabeth Dwoskin. 2019. YouTube's arbitrary standards: Stars keep making money even after breaking the rules. https://www.washingtonpost.com/technology/2019/08/09/youtubes-arbitrary-standards-stars-keep-making-money-even-after-breaking-rules/

[29] Elizabeth Dwoskin and Craig Timberg. 2019. Facebook bans extremist leaders including Louis Farrakhan, Alex Jones, Milo Yiannopoulos for being 'dangerous'. https://www.washingtonpost.com/technology/2019/05/02/facebook-bans-extremist-leaders-including-louis-farrakhan-alex-jones-milo-yiannopoulos-being-dangerous/

[30] EFF. 2020. Section 230 of the Communications Decency Act. https://www.eff.org/issues/cda230

[31] Phillip George Efthimion, Scott Payne, and Nicholas Proferes. 2018. Supervised machine learning bot detection techniques to identify social twitter bots. *SMU Data Science Review* 1, 2 (2018), 5.

[32] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. (2019).

[33] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[34] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Twelfth International AAAI Conference on Web and Social Media*.
[35] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[36] Claire Goforth. 2020. Banned by PayPal and YouTube, this alt-right comedian is back on PayPal and YouTube. https://www.dailydot.com/debug/owen-benjamin-violating-internet-bans/
[37] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.
[38] Richard Hanania. 2019. It Isn't Your Imagination: Twitter Treats Conservatives More Harshly Than Liberals. https://quillette.com/2019/02/12/it-isnt-your-imagination-twitter-treats-conservatives-more-harshly-than-liberals/
[39] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[40] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing twitter users who engage in adversarial interactions against political candidates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[41] Yiqing Hua, Thomas Ristenpart, and Mor Naaman. 2020. Towards Measuring Adversarial Twitter Interactions against Candidates in the US Midterm Elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 272–282.
[42] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2020. Still out there: Modeling and identifying russian troll accounts on twitter. In *12th ACM Conference on Web Science*. 1–10.
[43] Infowars.com. 2019. Bombshell: Gov Official Confirms Link Between Vaccines and Autism. https://www.infowars.com/bombshell-gov-official-confirms-link-between-vaccines-and-autism/
[44] Infowars.com. 2019. Governments Admit Geoengineering/Weather Modification is Live - Was Dorian Manipulated? https://www.infowars.com/governments-admit-geoengineering-weather-modification-is-live-was-dorian-manipulated/
[45] Sue Curry Jansen and Brian Martin. 2015. The Streisand effect and censorship backfire. (2015).
[46] Nash Jenkins. 2016. Twitter Suspends Conservative Writer Milo Yiannopoulos. https://time.com/4414400/milo-yiannopolous-twitter-abuse/
[47] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did You Suspect the Post Would Be Removed?": Understanding User Reactions to Content Removals on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 192 (Nov. 2019), 33 pages. https://doi.org/10.1145/3359294
[48] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Trans. Comput.-Hum. Interact.* 26, 5, Article 31 (July 2019), 35 pages. https://doi.org/10.1145/3338243
[49] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 150 (Nov. 2019), 27 pages. https://doi.org/10.1145/3359252
[50] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 12 (March 2018), 33 pages. https://doi.org/10.1145/3185593
[51] Jialun "Aaron" Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation Challenges in Voice-based Online Communities on Discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), Article 55. https://doi.org/10.1145/3359157
[52] Mücahit Kantepe and Murat Can Ganiz. 2017. Preprocessing framework for Twitter bot detection. In *2017 International Conference on Computer Science and Engineering (UBMK)*. IEEE, 630–634.
[53] Griffin Kelly. 2018. Twitter blocks Owen Benjamin. https://www.adirondackdailyenterprise.com/news/local-news/2018/04/twitter-blocks-owen-benjamin/
[54] Sara Kiesler, Robert Kraut, and Paul Resnick. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* (2012).
[55] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.
[56] Jason Koebler. 2018. Social Media Bans Actually Work. https://www.vice.com/en_us/article/bjbp9d/do-social-media-bans-work
[57] Rachel Kraus. 2018. 2018 was the year we (sort of) cleaned up the internet. https://mashable.com/article/deplatforming-alex-jones-2018/

[58] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3739–3748.
[59] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2004).
[60] Sébastien Liarte. 2013. *Brand image and Internet: Understanding, avoiding and managing the "Streisand" effect.* Technical Report. HAL.
[61] Andrew Marantz. 2020. Alex Jones's Bogus Coronavirus Cures. https://www.newyorker.com/magazine/2020/04/06/alex-jones-bogus-coronavirus-cures
[62] Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*. IEEE, 1–6.
[63] Alice Marwick and Rebecca Lewis. 2017. Media manipulation and disinformation online. *New York: Data & Society Research Institute* (2017).
[64] Adrienne Massanari. 2015. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* (2015). http://nms.sagepub.com/content/early/2015/10/07/1461444815608807.abstract
[65] Nathan J. Matias. 2016. The Civic Labor of Online Moderators. In *Internet Politics and Policy conference* (Oxford, United Kingdom). Oxford, United Kingdom.
[66] Stephanie Mencimer. 2016. PizzaGate shooter read Alex Jones. Here are some other fans who perpetrated violent acts. https://www.motherjones.com/politics/2016/12/comet-pizza-suspect-shooters-alex-jones/
[67] Alexandros Mittos, Savvas Zannettou, Jeremy Blackburn, and Emiliano De Cristofaro. 2019. "And We Will Fight For Our Race!" A Measurement Study of Genetic Testing Conversations on Reddit and 4chan. ICWSM (2019). arXiv:1901.09735 http://arxiv.org/abs/1901.09735
[68] Joe Mulhall. 2019. Deplatforming Works: Let's Get On With It. https://www.hopenothate.org.uk/2019/10/04/deplatforming-works-lets-get-on-with-it/
[69] Angela Nagle. 2017. *Kill all normies: Online culture wars from 4chan and Tumblr to Trump and the alt-right.* John Hunt Publishing.
[70] Adewale Obadimu, Esther Mead, Muhammad Nihal Hussain, and Nitin Agarwal. 2019. Identifying toxicity within youtube video comment. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 214–223.
[71] Abby Ohlheiser. 2019. Analysis | The 96 hours that brought down Milo Yiannopoulos. https://www.washingtonpost.com/news/the-intersect/wp/2017/02/21/the-96-hours-that-brought-down-milo-yiannopoulos/
[72] Zachary Petrizzo. 2019. Owen Benjamin, alt-right comedian, banned from YouTube. https://www.dailydot.com/layer8/owen-benjamin-youtube/
[73] Ashwin Rajadesingan, Paul Resnick, and Ceren Budak. 2020. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 SE - Full Papers (2020). https://www.aaai.org/ojs/index.php/ICWSM/article/view/7323
[74] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Robert West. 2021. Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels. (2021).
[75] Sarah T Roberts. 2017. *Content moderation.*
[76] Sarah T Roberts. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media.* Yale University Press.
[77] Richard Rogers. 2020. Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media. *European Journal of Communication* (2020), 0267323120922066.
[78] Aja Romano. 2016. Milo Yiannopoulos's Twitter ban, explained. https://www.vox.com/2016/7/20/12226070/milo-yiannopoulus-twitter-ban-explained
[79] Tony Romm. 2018. Twitter has permanently banned Alex Jones and Infowars. https://www.washingtonpost.com/technology/2018/09/06/twitter-has-permanently-banned-alex-jones-infowars/
[80] Tony Romm and Elizabeth Dwoskin. 2019. Twitter adds labels for tweets that break its rules - a move with potentially stark implications for Trump's account. https://www.washingtonpost.com/technology/2019/06/27/twitter-adds-labels-tweets-that-break-its-rules-putting-president-trump-companys-crosshairs/
[81] Haji Mohammad Saleem and Derek Ruths. 2018. The Aftermath of Disbanding an Online Hateful Community. *arXiv preprint arXiv:1804.07354* (2018).
[82] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.
[83] Ralph Schroeder. 2018. Rethinking digital media and political change. *Convergence* 24, 2 (2018), 168–183.

[84] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[85] Avi Selk. 2019. How deplatforming became a rallying cry for right-wing media stars. https://www.washingtonpost.com/lifestyle/style/how-deplatforming-became-a-rallying-cry-for-right-wing-media-stars/2019/07/10/f2f37a72-a348-11e9-bd56-eac6bb02d01d_story.html

[86] Olivia Solon. 2020. Sensitive to claims of bias, Facebook relaxed misinformation rules for conservative pages. https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182

[87] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content Removal as a Moderation Strategy: Compliance and Other Outcomes in the ChangeMyView Community. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 163 (Nov. 2019), 21 pages. https://doi.org/10.1145/3359265

[88] Ryan Suppe and Charles Ventura. 2018. Before Twitter suspended Alex Jones, Infowars was already directing users to Tumblr. https://www.usatoday.com/story/tech/talkingtech/2018/08/15/before-twitter-suspended-alex-jones-infowars-already-directing-users-tumblr/992263002/

[89] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018), 385–400.

[90] Nicolas P Suzor, Sarah Myers West, Tarleton Gillespie, and Jillian York. 2017. Guiding principles for the future of content moderation. (2017).

[91] **Shagun Jhaver***, Eshwar Chandrasekharan*, Amy Bruckman, and Eric Gilbert. 2021. Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. (2021).

[92] Daniel Trotta. 2019. Infowars founder who claimed Sandy Hook shooting was a hoax ordered to pay $100,000. https://www.reuters.com/article/us-texas-lawsuit-alex-jones-idUSKBN1YZ1BB

[93] Twitter. 2019. https://transparency.twitter.com/en/reports/information-operations.html. https://blog.twitter.com/en_us/topics/company/2019/18_midterm_review.html

[94] Aleksandra Urman and Stefan Katz. 2020. What they do in the shadows: examining the far-right networks on Telegram. *Information, Communication & Society* (2020), 1–20.

[95] Jonathan Vanian. 2017. Twitter Toughens Rules on Nudity and Revenge Porn | Fortune. http://fortune.com/2017/10/27/nudity-revenge-porn-twitter/

[96] Mark Wilson. 2019. Pinterest is escalating its fight against anti-vaxxers as measles surge in the U.S. https://www.fastcompany.com/90396075/pinterest-is-escalating-its-fight-against-anti-vaxxers-as-measles-surge-in-the-u-s

[97] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 160.

[98] Graeme Wood. 2017. How Richard Spencer Became an Icon for White Supremacists. https://www.theatlantic.com/magazine/archive/2017/06/his-kampf/524505/

[99] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1391–1399. https://doi.org/10.1145/3038912.3052591

[100] Milo Yiannopoulos. 2014. Feminist Bullies Tearing the Video Game Industry Apart. https://www.breitbart.com/europe/2014/09/01/lying-greedy-promiscuous-feminist-bullies-are-tearing-the-video-game-industry-apart/

[101] Milo Yiannopoulos. 2014. GamerGate: Angry Feminists, Unethical Journalists Are the Ones Not Welcome in the Gaming Community. https://www.breitbart.com/entertainment/2014/09/15/the-gamergate-movement-is-making-terrific-progress-don-t-stop-now/

[102] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. 188–202.

[103] Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. (2020), 125–134. https://doi.org/10.1145/3394231.3397902 arXiv:2005.07926

A IDENTIFYING ALEX JONES SUPPORTERS

As noted in section 3.4, the Alex Jones dataset contained both the supporters and opponents of Alex Jones. We describe here how we identified the Alex Jones supporters. Note that this analysis includes only those users who posted at least 10 times pre-deplatforming of Alex Jones.

We observed that the most frequently used hashtags in this dataset included "*#FreeAlexJones*" and "*#BanAlexJones*." Our manual analyses of 50 tweets containing each of these keywords also showed that the posters clearly displayed a pro-Alex Jones stance when using "#FreeAlexJones" and anti-Alex Jones stance when using "#BanAlexJones." For instance, these tweets are typical examples of posts using the keywords "#FreeAlexJones" and "#BanAlexJones" respectively:

> *"Infowars is the least biased of all major news sources! #StopTheBias #FreeAlexJones #ImAlexJones #HillaryForPrison"*
>
> *@Twitter @TwitterSupport @TwitterSafety you should ban Alex Jones! #BanAlexJones*

In view of this, we collected users who used the hashtag "#FreeAlexJones" at least 5 times but never used the hashtag "#BanAlexJones" in the Alex Jones dataset, and labeled them as *supporters*. Similarly, we collected all users who employed the hashtag "#BanAlexJones" at least 5 times but never used the hashtag "#FreeAlexJones," and labeled them as *opponents*. Through this process, we labeled 602 users as *supporters* and 110 users as *opponents*.

To evaluate the quality of these labels, we randomly sampled 50 *supporters* and 50 *opponents*, and manually reviewed their tweets to check their stance towards Alex Jones. This manual review did not change our labels for any user, which indicated that our approach to identify supporters and opponents had a high precision.

Next, we used the labeled supporters and opponents to build machine learning classifiers and conduct label propagation. We first collected tweets posted by *supporters* and *opponents* in the α-D dataset for Alex Jones. We used many features to build these classifiers that we describe next:

A.1 Classifier Features

We used three sets of features: (1) Behavioral features, (2) Hashtag features, and (3) Bag of words features. Next, we detail these features.

Behavioral Features: (1) Number of tweets, (2) Average number of hashtags per tweet, and (3) Average number of mentions per tweet. We used these features because our manual analysis of tweets posted by *supporters* and *opponents* indicated that Alex Jones supporters were more active than opponents, and they often used many hashtags and mentions in their tweets as a way to reach broader audiences. For example, one *supporter* posted this tweet:

> *@prisonplanet @randpaul @realalexjones #texas #dallas #garland #tcot #ttcot #wheelsup indicting corrupt public official w grand jury @donaidtrumpreal how to clean out the swamp in your own back yard... https://www.real.video/58270767930010*

Hashtag Features: We extracted each hashtag used by *supporters* and *opponents*, and constructed a vocabulary containing only the top 5000 most frequent hashtags. We used these 5000 hashtags as features and calculated their values to be the relative frequency of occurrence of the corresponding hashtag in their tweets. We did not use the hashtags "#FreeAlexJones" and "#BanAlexJones" as features because they were used to assign the initial labels.

Bag of Words Features: Finally, we tokenized the tweets (converted to lowercase), i.e., we divided the tweet text into words, and discarded stopwords.[16] Using these words, we extracted unigrams, bigrams and trigrams from the text, and constructed a vocabulary containing only the

[16]We used the English language stopwords derived from NLTK corpus. We also added "#FreeAlexJones" and "#BanAlexJones" as stopwords.

5000 most frequent *n*-grams. Next, we extracted the term frequency-inverse document frequency (TF-IDF) values of these *n*-grams for each user [2]. The TF-IDF algorithm weighs how often a term appears in a document (tweet), but normalizes this weighting by the frequency of that term in the entire collection. We used the TF-IDF values derived through this process as features during the classification phase.

A.2 Classification Tests

We ran classification tests using Logistic Regression, Random Forest, Support Vector Machine, and Naive Bayes algorithms, using all three categories of features described above and optimizing each classifier's performance through a parameter search. These tests showed the best F-1 scores when using Logistic Regression with liblinear algorithm and L1 regularization. A 10-fold cross validation test using this algorithm gave the following mean performance:

Precision: 0.983
Recall: 0.995
F-1: 0.989

We next used this best-performing classifier to label the rest of users in the Alex Jones dataset. To ensure a high accuracy of our labels, we sorted users classified as supporters (opponents) in decreasing order of the classifier's probability estimates, and assigned the top third of such users as *supporters* (*opponents*).

A.3 Validation through human evaluation

As an additional validation step, we employed human evaluation to assess the quality of classifier predictions. We first randomly selected 50 users labeled as *supporters* and 50 users labeled as *opponents* by our classifier. Next, we randomly selected 15 tweets that each user posted. A coder manually reviewed each user's tweets and assigned that user an "Alex Jones supporter" or "Alex Jones opponent" label. This post-hoc analysis showed that our classifier labels matched with the coder's labels in 94 out of 100 cases. We considered this performance as sufficiently accurate to conduct our subsequent data collection and analyses.