Journal of Quantitative Description: Digital Media 1(2023), 1–44 DOI 10.51685/jqd.2023.004

Cross-Platform Reactions to the Post-January 6 Deplatforming

CODY BUNTAIN
University of Maryland, USA

MARTIN INNES
Cardiff University, UK

TAMAR MITTS
Columbia University, USA

JACOB N. SHAPIRO
Princeton University, USA

We study changes in social media usage following the 'Great Deplatforming' in the aftermath of the 6 January 2021 attack on the US Capitol. Following the attack, several major platforms banned thousands of accounts, ostensibly to limit misinformation about voter fraud and suppress calls for violence. At the same time, alternative platforms like Gab, BitChute, and Parler welcomed these deplatformed individuals. We identify three key patterns: First, in studying the platforms that emerged among users seeking alternative spaces, we see high frequencies of users bridging these communities announcing their intent to join non-mainstream platforms to their audiences on mainstream platforms. Second, focusing on platforms that were created to be alternative, anti-censorship spaces, deplatforming preceded a sustained increase in engagement with Gab across Twitter, Reddit, and Google search, while Parler saw a steep decline in engagement. Third, examining the language in these spaces, toxic discourse increased briefly on Reddit and Twitter but returned to normal after the deplatforming, while Gab became more toxic. These results suggest that while deplatforming may precede a reduction in targeted discussions within a specific platform, it can incentivize users to seek alternative platforms where these discussions are less regulated and often more extreme.

Keywords: Deplatforming, Social Media, Alt Tech Platforms

Cody Buntain (Corresponding Author): cody@bunta.in
Martin Innes: innesm@cardiff.ac.uk
Tamar Mitts: tm2630@columbia.edu
Jacob N. Shapiro: jns@princeton.edu
Date submitted: 2022-12-13. This version 2022-03-02.

Introduction

Large parts of the world communicate on social media. Nearly a third of the world's population has Facebook accounts. As more people move online, individuals inclined to racism, vulgarity, misogyny, or homophobia have found niches where they can express their views in ways that magnify discord and violate platforms' rules for appropriate behavior. For instance, the British far-right political commentator, Milo Yiannopoulos, gathered a large cohort of anonymous [online] activists on Twitter who magnified his calls for targeted harassment (Jenkins, 2016). Another influencer, Alex Jones, marshaled thousands of followers on social media to promote his conspiracies, which led to violent acts (Mencimer, 2016). And an alt-right former comedian, Benjamin Owen, garnered supporters by repeatedly making anti-Semitic and bigoted statements, as well as spreading misinformation (Goforth, 2020). Perhaps the most prominent case is that of former U.S. President Donald Trump, whose persistent tweets about voter fraud after the 2020 U.S. elections encouraged crowds to march to the U.S. Capitol on January 6, 2021 (see e.g. Subramanian, 2021).

One response to such online conversations is identifying and banning the accounts driving them – a process known as deplatforming. As a policy tool, deplatforming intends to reduce misinformation and extremism by expelling malicious content creators from social media platforms, deterring future offensive behaviors, and reducing the overall toxicity of speech online (Jhaver et al., 2021).

A few academic studies suggest that deplatforming is indeed an effective tool for reducing the impact of malign actors on the public (Jhaver et al., 2021; Rauchfleisch and Kaiser, 2021). At the same time, other studies, such as Rohlinger et al. (2023), contradict this apparent efficacy, finding instead that deplatforming is ineffective at improving the information space. Even in the case of a reduction in the amount of toxic content on one platform after deplatforming, however, this behavior might not be a sign of reduced impact, as studies contend. After being removed from social media sites, banned actors can migrate to less regulated platforms, potentially promoting even more radical ideas (Urman and Katz, 2022; Bryanov et al., 2022). In addition, the set of followers on alt-social sites might be individuals with ideologies closer to that of banned content creators. In such cases, deplatforming individuals is an incomplete way to curb harmful discourse. It could exacer-

bate misinformation and extremism by pruning extraneous followership and catalyzing the aggregation of ideologically similar people in alt-social sites (Rogers, 2020; Coaston, 2018). Deplatforming can also bring attention to the 'victimized' persons (Ohlheiser, 2016), which can motivate toxic content creators to move to alternate platforms where they can continue engaging and reach followers with fewer constraints.

We contribute to the growing literature on deplatforming by studying trends in online activity on mainstream and alt-social media sites around the January 6 attack on the U.S. Capitol, which motivated large, mainstream platforms to remove dozens of prominent individuals and suspend thousands of accounts from their sites. We examine three aspects of online activity. First, we provide qualitative evidence on the way in which account holders on mainstream platforms talked about joining other, less-mainstream social media sites. Second, we assess how engagement with a set of explicitly anti-censorship alternative sites on mainstream platforms changed. Finally, we use event studies to understand how discourse shifted on Reddit, Twitter, and Gab–a platform that gained a significant number of users following the deplatforming. These analyses do not isolate the causal impact of removal, however, as many aspects of American political discourse shifted after January 6, and the media coverage of this deplatforming was intense and polarized. Instead, this work provides descriptive evidence that the deplatforming correlated with major changes in social media engagement that should induce caution and require further study.

Our data shows that much of the engagement with alternative platforms was announced on Twitter and Facebook. Among alt-social media platforms, Gab saw a sustained increase in attention on mainstream platforms. When examining how the discourse changed, we find that Gab became much more toxic after the deplatforming waves, with hate speech rising to levels that were much higher than previous months. Other platforms saw a more modest change in toxic content. On Twitter, for example, hate speech spiked dramatically in the week following January 6 and then fell to levels that were 10-15% above the month of December in the following 2 months. On Reddit, the changes were quite minimal, with only some types of hate speech spiking after January 6.

This work breaks new ground by looking at both movement and change in followership as actors jump platforms and how the nature of the discourse changed over an

extended period. Our results are consistent with a dynamic in which deplatforming drives local improvements that are potentially at the expense of the larger social ecosystem. Our description of response to this deplatforming also lays groundwork for where deplatforming may be effective or not more generally.

To lay the foundation for this work, we next provide background on "The Great Deplatforming" and the lack of consensus in the literature around the effects of deplatforming on audiences and spaces. We then lay out three areas of study to describe user- and platform-level behaviors in the lead-up to, during, and after The Great Deplatforming. First, we present an exploratory analysis of social media users' emergent interests in non-mainstream social media spaces, identifying several platforms–both explicitly anti-censorship spaces, like Gab/Parler, and privacy-oriented platforms, like Telegram and MeWe–that saw increased engagement following The Great Deplatforming. Second, we move from describing platforms that emerge as alternative spaces in users' discourse to describing dynamics among platforms that explicitly style themselves as alternative, anti-censorship spaces. Finally, we examine how the nature of conversations on various platforms changed before and after January 6, by analyzing trends over time in toxic content. Building on these analyses, we then discuss implications for policy, how these results might generalize to other deplatforming interventions, and areas of future research, including a potential spectrum of deplatforming interventions.

Background

January 6 and the Great Deplatforming

On January 6, 2021, thousands of President Trump's supporters stormed the U.S. Capitol building in an attempt to overturn the 2020 presidential election. The rally was not spontaneous, nor was the march on the Capitol. Both were openly planned on social media platforms, including Twitter and Facebook, and both were encouraged by elected government officials (Polantz et al., 2021; Fuchs, 2021). Before the event, President Trump tweeted intensively about voter fraud and encouraged the crowd to march to the Capitol (see e.g. Subramanian, 2021).

Immediately after the riot, Twitter blocked President Trump from posting, and

deleted the tweets he sent during the Capitol Attack. Two days later, on January 8, the President's personal account @realDonaldTrump was permanently suspended (Citron, 2021; Mak, 2021b; Twitter, Inc., 2021). Facebook also banned President Trump from using Facebook and Instagram until at least Inauguration Day (Mak, 2021a). On January 10, Amazon stopped providing web hosting services to Parler, a social media network where Trump supporters were rallying together and praising the Capitol Attack (Romm and Lerman, 2021). Soon after that, YouTube also suspended Trump's channel for inciting violence (Mickle, 2021), and TikTok removed videos of Trump's speeches and blocked hashtags related to the Capitol riot (Elegant, 2021).

President Trump was not the only person removed from mainstream platforms following the attempted insurrection. After the event, Facebook removed over 20,000 groups and pages linked to the event, and Twitter suspended more than 70,000 accounts linked to the Capitol attack, claims of voter fraud, and QAnon (Conger, 2021; Sullivan, 2021). YouTube began taking down channels owned by high-profile election-fraud activists in 2021, though they have not disclosed specific details on the number of channels removed (De Vynck, 2021). Collectively, these actions amounted to an unprecedented coordinated response to the threat posed by an online community.

Deplatforming: What We Currently Know

Research on the impact of deplatforming has not arrived at a consensus. Below, we summarize evidence for and against deplatforming as a policy tool.

Deplatforming Reduces Misinformation and Extremism

The few academic studies on the topic so far suggest that deplatforming is an effective tool for expelling malicious content creators from mainstream platforms, deterring future offensive behaviors, and reducing the overall toxicity of online speech. Jhaver et al. (2021) examined the effects of Twitter's deplatforming of Alex Jones, Milo Yiannopoulos, and Owen Benjamin from Twitter, three prominent misinformation purveyors. The authors found that, after their removal, conversations about their accounts declined significantly on

Twitter, as did the general toxicity of their supporters' posts.

In a similar vein, Chandrasekharan et al. (2017a) examine the effectiveness of deplatforming on Reddit. Analyzing hate speech generated around the 2015 ban of the subreddits r/fatpeoplehate and r/CoonTown, the authors found that users who participated in these communities either significantly decreased their use of hate speech, or left the platform altogether. This research also looked into spillover effects and did not find significant changes in the tone of other subreddits.

Analyzing spillover effects more closely, Rauchfleisch and Kaiser (2021) tracked the activity of more than 10,000 YouTube channels removed between January 2018 and October 2019, and checked whether similar accounts were opened at BitChute (a YouTube clone known to include significant political and hateful content Trujillo et al. (2020)). While similar accounts were opened, the alternative platform reached a much smaller audience and thus, they conclude, deplatforming is effective in minimizing the spread of malicious content in general. Along similar lines Seering et al. (2017) show that proactive regulation, such as restricting posting of certain kinds of content, can be effective in discouraging spamming. Users in the chatroom observing other users being banned engaged in less hostile speech than before the restrictions were put in place.

Buntain et al. (2021) further studied YouTube's recommendation-driven deplatforming strategy in 2019, wherein YouTube would no longer show videos deemed "potentially harmful or misinforming" in recommended feeds or "Up-Next" automatic playlists – though these videos would remain on the platform and accessible via their parent channels, direct links, or search. Results from Buntain et al. (2021) demonstrated that engagement with classes of potentially treated/de-recommended videos on YouTube decreased significantly following YouTube's de-recommendation announcement. These results were consistent across Twitter and Reddit, suggesting recommendation-oriented deplatforming – potentially reducing monetary incentives to produce such content – appeared to have a suppressive effect on the larger information ecosystem.

More anecdotal evidence echos the findings of these studies. Nouri et al. (2019) studied the removal of "Britain First," a UK far-right group with 1.8 million followers, from

Twitter in December 2017 and Facebook in March 2018. That removal appeared to have diverted the group's followers to smaller platforms such as Gab and Telegram. The group went from 1.8 million Facebook followers to a mere 12,000 followers on Gab. The report's authors take this as a sign of reduced impact.

Deplatforming Exacerbates Misinformation and Extremism

Other studies argue that deplatforming can drive extremist speech to darker corners, bring attention to censored persons, and reinforce their identity and radicalize ideology. Ali et al. (2021) developed a method for following accounts between platforms and estimated that the majority of the suspended accounts on Twitter (58.74%) and Reddit (75.88%) created accounts on Gab, where most of them spread far-right ideas. Mitts (2022) found similar patterns when examining the reactions of Gab users when their accounts were suspended from Twitter. Broadly speaking, there is good evidence that creators of fringe content turn to Gab (Zhou et al., 2019; McIlroy-Young and Anderson, 2019), Parler (Munn, 2021), BitChute (Trujillo et al., 2020), Discord, Telegram (Rogers, 2020), 4chan (Bernstein et al., 2011), and other small platforms (e.g., Something Awful Forums, see Pater et al., 2014) and even the dark web (Robertson, 2017), when removed from mainstream social media.

If suspended users systematically move to fringe platforms, then a critical question is whether suspended account-holders can still reach a substantial audience. Cofnas (2019) argue that censoring can backfire as the action of banning itself attracts attention to the 'victimized' person and their alternative platforms. Likewise, content moderation studies have shown that perceptions of moderation actions (e.g., whether they are "unfair" or transparent in their application) drive recidivism and anti-social behavior (Cheng et al., 2015; Chang and Danescu-Niculescu-Mizil, 2019). After Facebook announced it would remove Yiannopoulos, Jones, Laura Loomer and Paul Joseph Watson in 2019, these account holders pointed their followers to alternative platforms (Martineau, 2019). And some far-right YouTube pundits have built similar audiences on BitChute as they had on YouTube (Rauchfleisch and Kaiser, 2021, p. 6). If removed users can attract similar numbers of "true believers" on alternative sites, then perhaps deplatforming will not actually reduce radicalization.

A related concern is that deplatforming will elevate the level of extremism in general, as banning users from mainstream social media only makes them more alienated (in line with past content moderation work), less likely to change their minds, and thus reinforces their extreme ideas (Cofnas, 2019). Quantitative studies on community feedback do confirm the detrimental potential of "stigmatizing" users, which can drive them to post more frequently with lower quality (Cheng et al., 2014, 2015; Chang and Danescu-Niculescu-Mizil, 2019). In the most direct tests of this concern Mitts (2022), Ribeiro et al. (2020), and Ali et al. (2021) found that, after migrating to alternative platforms, suspended account holders produced a larger volume of more toxic content and hate speech.

Trends in User Interest in Alternative Platforms after January 6

To examine how affected users expecting to be deplatformed tried to move their communication to alternative spaces after the Great Deplatforming, we draw on several sources of data, including Brandwatch, Google Trends, and Meta's Crowdtangle tool. We describe time trends in social media users' interest in alternative platforms, highlight those that gained the most attention, and discuss how these users talked about the move. In this section, we focus on the platforms which were mentioned organically at the time, some of which have since faded in prominence.

Figure 1 shows time trends in Twitter mentions of Parler, Gab, Telegram and Mewe between October 2020 and October 2021, based on data obtained from the Brandwatch tool. While mentions of these alternative platforms do not necessarily indicate an interest in migration, they can serve as a proxy for users' levels of attention to these sites. We find that in November 2020, around the time of the U.S. Presidential election, Parler experienced a growth in interest. In early January, Parler, Gab and Telegram were mentioned more frequently on Twitter, with Parler's growth being far in excess of that detected for any of the others. But in the longer term, most of these platforms did not sustain interest on Twitter, other than Telegram. The surge in user mentions of Telegram in early January turned into a sustained growth in the following months – suggesting that Telegram may have become an increasingly important platform in the media ecosystem.

Figure 2 shows the temporal trends in weekly mentions of 'switching' to each of the

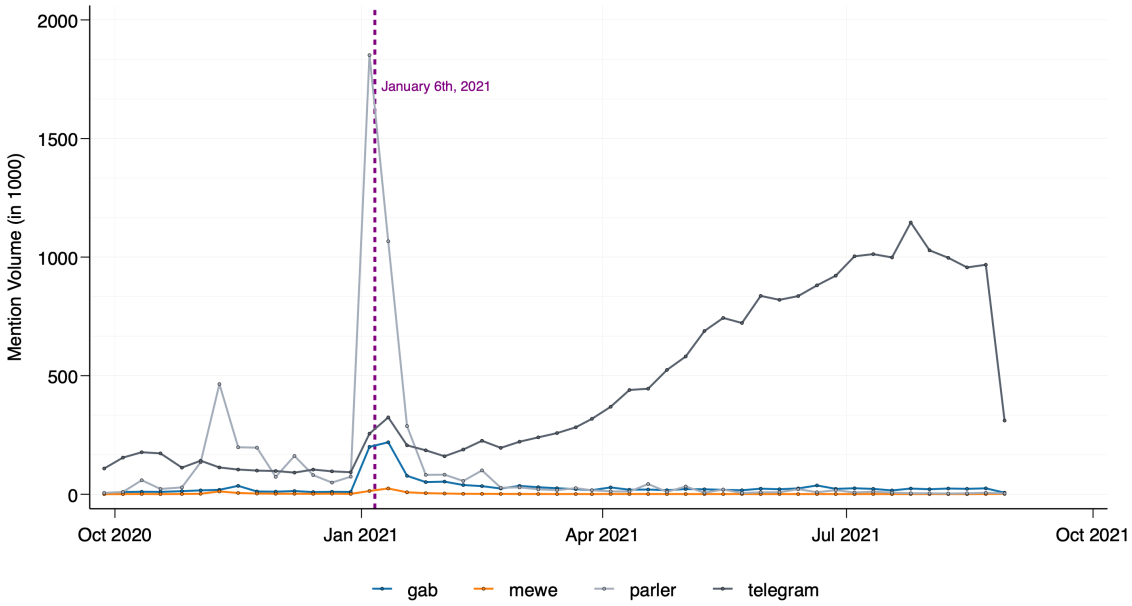


Figure 1. Platform Mentions on Twitter

Note. Figure plots weekly mentions of alternative platforms on Twitter. Data from Brandwatch in October 2021.

destination platforms in turn, based upon Twitter mentions, including where tweets have used a URL shortener. We see that:

- Parler saw a period of rapid growth in November, and then a dramatic spike in January. However, the sharing of links virtually ceased once the Parler.com website was taken offline.
- Gab experienced a smaller growth in interest in November, but its popularity surged in January as Parler waned. Interest has since dropped, but has remained above its pre-January levels.
- MeWe also had a peak in November and a large peak in January, but this interest did not last in the longer term.
- Telegram experienced a 'slow burn' growth inasmuch as it really started to take off

in March and has surged in popularity since (we cannot explain the recent reduction in the Brandwatch data).

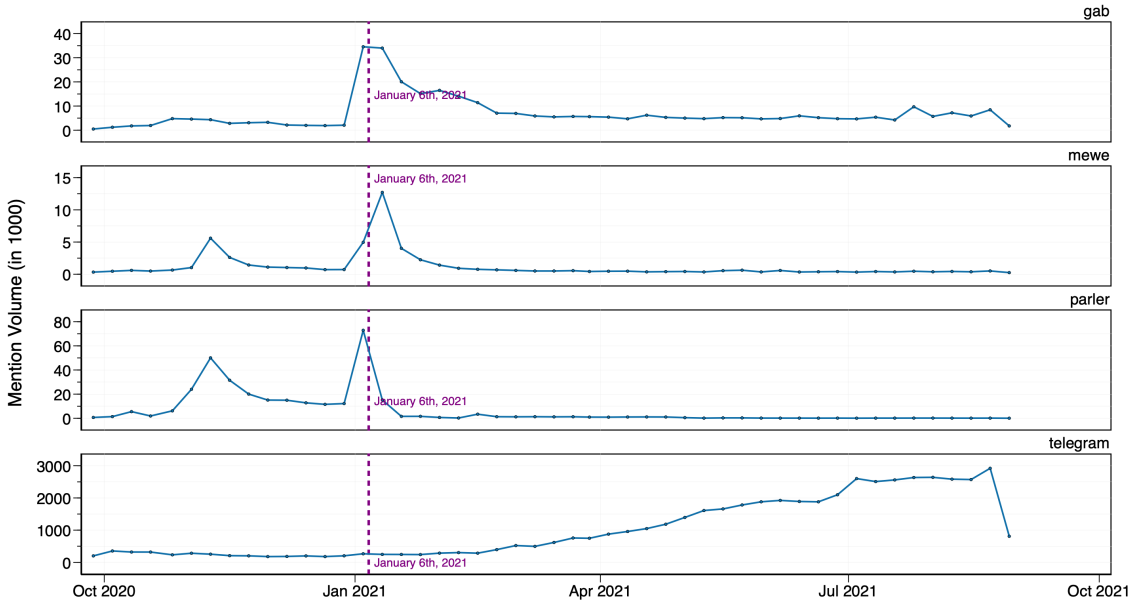


Figure 2. Switching Discourse Regarding Different Platforms

Note. Figures plots weekly tweets containing links to Parler, Gab, MeWe, and Telegram. Data from Brandwatch in October 2021.

To examine whether these trends exist in other sources of data, we use Google Search trends to complement the evidence above. Google Trends allows researchers to access information on search terms over time. In Google Trends, the data is normalized to lie between 1 and 100 for a given sample period, with the highest number during the sample period being given a score of 100. Table 1 shows Google search interest in these four alternative platforms from October 2020 through April 2021 where the origin of the search is in the United States. Since the scores are not directly comparable, we compared each score given in January against the search term 'Twitter', showing the score that the destination platform keyword received relative to Twitter in parentheses. The results presented in the table show that there was a spike in interest in Gab, Parler, and Telegram in November 2020, echoing our Brandwatch findings. We also find that Parler and MeWe experienced

peaks in search interest during the U.S. election period, whereas Gab and Telegram saw a growth in interest a bit later, in January.

Table 1: Google Trends audience interest for four destination platforms

Date	Parler	Gab	Telegram	MeWe
October - 2020	2	2	28	3
November - 2020	78	3	31	100
December - 2020	3	3	32	7
January - 2021	100(47)	100(8)	100(3)	72(3)
February - 2021	8	19	44	6
March - 2021	1	7	42	3
April - 2021	1	6	37	3

To provide an intuition for the nature of users' interest in migration, we next present qualitative data on Facebook users' discourse on Rumble, an alt-social platform that was mentioned frequently on Facebook during the period of interest. Using Meta's CrowdTangle tool, we collected data on 234,320 posts that were shared on public pages and groups on Facebook between October 1, 2020 and April 30, 2021 that mentioned the keyword 'Rumble.' One key disadvantage of Facebook data collected through Crowdtangle is that it does not include activity by accounts that were taken down before the date of our search.[1] The CrowdTangle data discussed below is thus an undercount of the total volume of intentional direction to Rumble which took place on Facebook.

Users posting about migration to Rumble typically explained or justified their move, discussed the benefits of migration, and provided details on where to find them on the alternative site. The clear intent was to persuade their Facebook followers to come follow them on Rumble. The following was typical:

> RUMBLE Rocks! Ben and I just launched Ultimate Survival Tips on RUMBLE. What's RUMBLE? Well it's a Social Media sight [sic] similar to YouTube

[1]Relatedly, a search today on the same keyword would return different results as any accounts or posts removed after May 2021. Note also that CrowdTangle does not provide access to posts from private user accounts or private Groups.

> but without all the shadow banning, deplatforming and politically correct nonsense. Check out the new review we just posted there. And don't forget to subscribe. Talk soon!!! -David ... https://rumble.com/vfb0mp-new-sol-waterproof-fire-lite-fuel-free-plasma-survival-lighter-review.html

A second example illustrates how users saw deplatforming as an unjustified censorship tool, and sought, through their Facebook posting, to invoke a sense that their move reflects a reaction to long-term trends in mainstream platforms' content management:

> Hello Patriots. I hope everyone is well, and healthy. With all of the censorship on the mainstream platforms, a lot of the great Patriot channels and accounts are getting banned and suspended. X22 Report was my go to for years on YouTube. Very informative, and thats why they censored him. Now you can find Dave and his program on Rumble. https://rumble.com/c/X22Report

For our purposes, the most interesting posts mentioning Rumble were those that were political in nature. Thus, to distinguish between political posts and those that were focused on other issues, we labeled posts 'political' if the Message, Image, Link Text or Description fields contained any of the following keywords: 'Trump,' 'Biden,' 'MAGA' or 'Capitol.'

Figure 3 shows the time trends in the number of 'political' and 'other' posts mentioning Rumble each day. We find that political posts spiked right before the election at the end of October, and then dropped, before rising again in the middle of December. The peak in Rumble posts of political nature peaked on the 13th January, when over 1,100 posts mentioning Rumble contained at least one of our 'political' keywords. In the following months, there was an extended period of decline, with the number of posts reducing to roughly 200 per day by the end of April. This analysis suggests that Rumble was a significant destination of interest for Facebook users and their followers during this period.

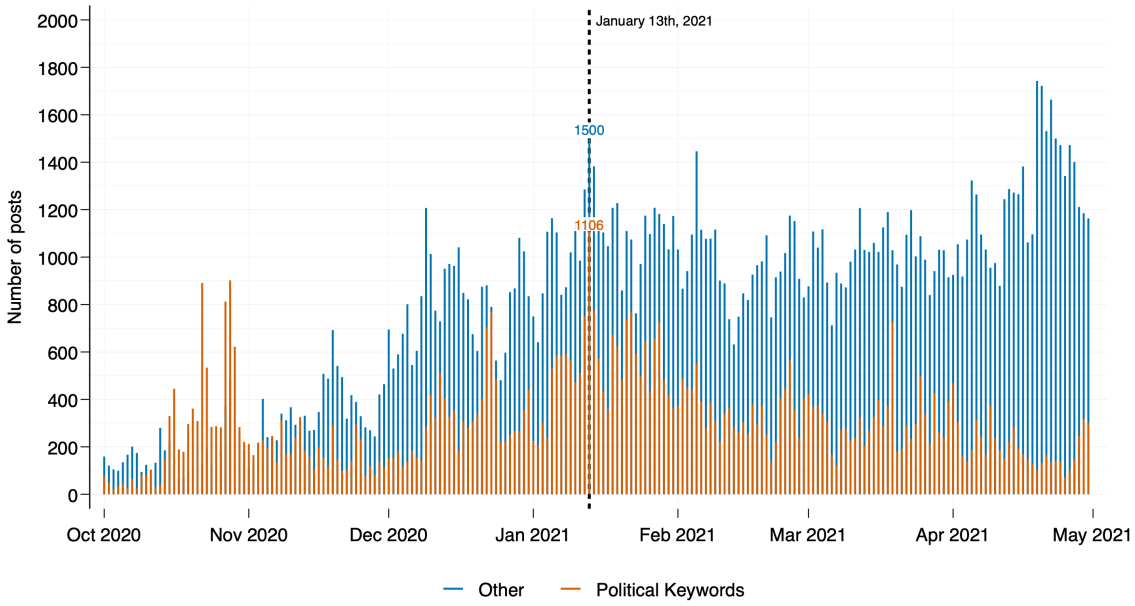


Figure 3. Public Facebook page and group political posts linking to Rumble

Note. Time series for mentions of Rumble on Facebook in political vs. non-political posts. Data from CrowdTangle.

Trends in Hyperlinking to Avowed Alt-Social Sites

We next turn to a descriptive analysis of engagement with three platforms whose purpose is to provide safe harbors for content and creators who have been removed from or made unwelcome by the mainstream platforms, namely BitChute, Parler, and Gab. While spaces like Rumble, Telegram, and MeWe are "alternative" in that they offer different affordances and priorities than their mainstream analogs, their core motivations are neither politically focused nor built on providing censorship-free spaces like BitChute, Parler, and Gab. This section therefore focuses on the supply-side aspects of alternative platforms, studying platforms that were created to fill a perceived market need for 'free speech,' where creators who have been deplatformed from other spaces are welcomed and their con-

tent fostered.[2] To describe engagement with these alt-social spaces, we study hyperlinking behaviors–i.e., URL sharing–across Twitter and Reddit, which we then couple with search interest via Google Trends.

To get Twitter data, we leverage an archive of tweets collected from Twitter's decahose stream, a 10% random sample of all tweets, starting on 1 September 2020 and going until 31 May 2021 (representing four months before and four months after the insurrection and resulting deplatforming in January). This dataset contains 10,594,743,385 tweets over 273 days, for an average of 38.8 million tweets per day. Studies on this data source have identified shortcomings in its use for tracking topical coverage over time Morstatter et al. (2013), but it should be sufficient for gauging changes in popularity of individual links, as suggested in the stream mining chapter of Leskovec et al. (2014). To estimate trends pre- and post-deplatforming, we measure the daily frequency of these Twitter posts, as shown in Figure 4.

Figure 4 includes both post counts and daily volume of link sharing, as off-platform link-sharing is our primary variable of observation for cross-platform engagement. Links account for about 1/3 of content shared on Twitter, totaling 3,378,193,518 tweets with links, or 12 million tweets with links per day.

To get Reddit data, we used the PushShift.io collection (Baumgartner et al., 2020) to obtain all Reddit submissions posted during the same timeframe as the Twitter data (1 September 2020 to 31 May 2021). This resulted in 278,957,729 submissions over 273 days (about 1.02 million submissions per day). In Reddit, our focus is on submissions, to the exclusion of Reddit *comments* (messages posted in response to submissions). While comments are more numerous, Reddit's norms are such that submissions are the primary means for link-sharing on the platform, and comments are more textual discussion. Likewise, Reddit submissions are more comparable to posts on Facebook pages and groups.[3] Figure 5 shows counts of Reddit submissions and daily volume of link sharing. This volume totals 184,692,148 links shared (though many may be duplicates), averaging 677,000 per day.

[2]Truth Social is a similar politically motivated platform which would fit well in this analysis but did not exist during our study time frame. For a fuller account of the alternative social media

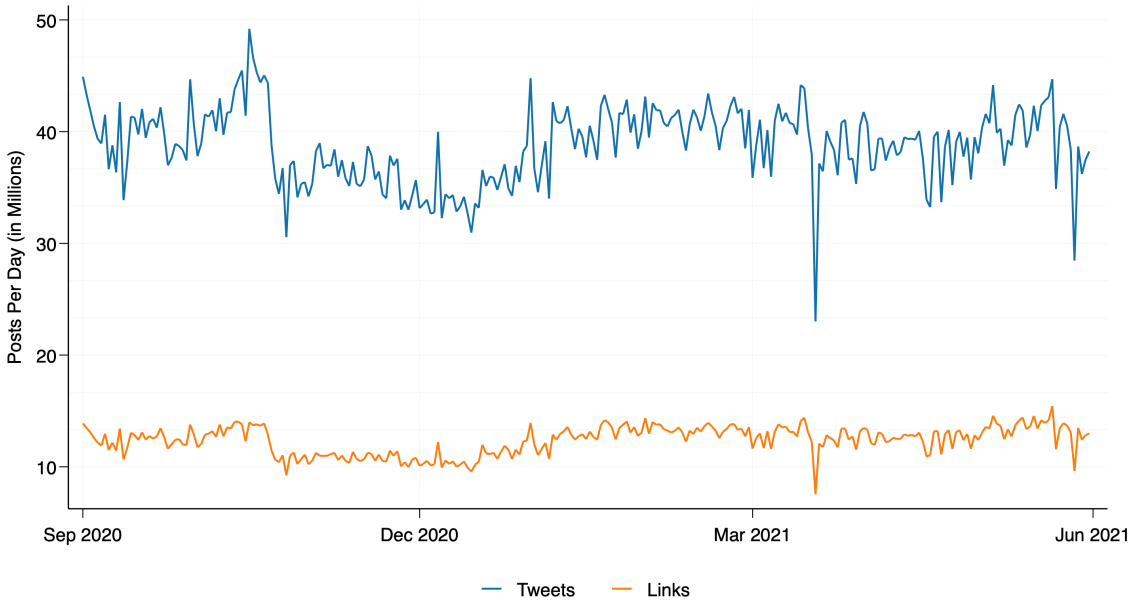


Figure 4. Time series for Twitter Data

Note. Data show general stability over the 9-month period, with a slight reduction between 23 October and 16 December. No clear change in behavior is apparent on 6 January. Data from Twitter decahose stream.

Quantifying the trajectory of engagement with alternative online spaces

To quantify changes in engagement trajectories, we use time series data extracted from the above sources and an interrupted time series (ITS) analysis (Bernal et al., 2018). The ITS model we leverage is similar to that employed in Chandrasekharan et al. (2017b) for measuring impact of banning in Reddit and Buntain et al. (2021) for evaluating cross-platform impact of de-recommendation in YouTube. While ITS is often used to assess efficacy of interventions, we use this method here only to describe changes in observed sharing behaviors. Many potential factors, from coordinated deplatforming to media coverage of deplatforming to President Trump's railing against moderation, all likely contribute to changes in these engagement behaviors, so we explicitly eschew any discussion of causal

environment, see Stocking et al. (2022).

[3]In subsequent analysis, we intend to run a robustness check using Reddit comments as well.

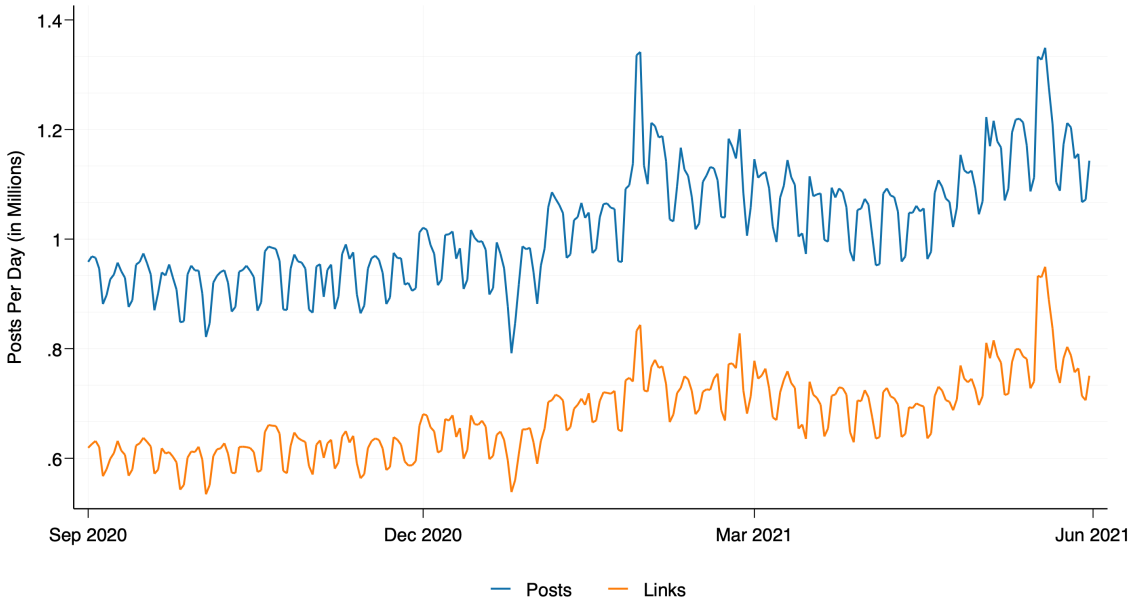


Figure 5. Reddit Trends

Note. Time series for Reddit, showing a general trend upward over the 9-month period, weekly periodicity, and peaks in activity on 29 January and 5 May. Data from PushShift.io.

mechanism here. Instead, we opt for ITS analysis as a construct for describing the behavioral changes we observe in these platforms, capturing both level (i.e., immediate) and trend (i.e., compounding effects as the treatment timeframe recedes in time) changes.

Our ITS model is defined in a general form in Equation 1. The trend factor here is of special interest because one might be less interested in the immediate changes following a media elite posting about an alternative platform compared to whether overall interest in that alternative platform is trending upward over time. Breaking down the model, it predicts the log-transformed number of shares to one of our three target platforms on a given day as a function of six factors:

1. the overall number of links shared on the host platforms l_t (whether the host platform is Twitter, Reddit, Google Trends, etc.),

2. how often links to this alternative platform p were shared on the previous day $l_{p,t-1}$ (where p is Gab, Parler, or BitChute),

3. news coverage of this platform on that day n_{t-1},

4. whether the day is during the "rollout period" $E(t)$ – i.e., 1 if this day is in the interim timeframe between 6 January and Twitter's banning of voter-fraud-related accounts on 12 January and 0 otherwise,

5. whether the day is before or after the deplatforming treatment $T(t)$ – 1 if so and 0 otherwise, and

6. how many days have passed since this treatment $d(t)$.

$$\ln(s_{p,t}+1) = \beta_1 \ln(l_t+1) + \beta_2 \ln(l_{p,t-1}+1) + \beta_3 \ln(n_{p,t}+1) + \beta_4 E(t) + \beta_5 T(t) + \beta_6 d(t) \quad (1)$$

The model uses log-transformed sharing volumes to account for the highly skewed nature of sharing in social networks, leading us to expect proportional changes in response rather than directly linear changes. Furthermore, we focus on the *percentage of links* shared on each platform rather than the percentage of *posts* because overall post volumes are subject to numerous additional factors (e.g., event responses). Instead, our focus is on the *distribution* of links to avoid confounders from overall social media activity. Then, for each mainstream-to-alternative-platform pair – {Twitter, Reddit, Google Trends, Facebook} X {Gab, Parler, BitChute, YouTube}, we use ordinary least-squares to fit the above model.

A core question around deplatforming is whether its application in mainstream platforms pushes audiences to more extreme online spaces. While we cannot narrow effect to deplatforming specifically versus its coverage or audience response to perceptions of deplatforming, we can describe changes in sharing behaviors. If deplatforming indeed pushes media elites and audiences to these more extreme spaces, one would expect to see a significant increase in the cross-platform sharing of these sites. Similarly, if deplatforming encourages audiences to migrate to these alternative sites, one would also expect to see a more general increase in searches for these platforms, which we measure via Google Trends. For comparison, we also include links to YouTube across these spaces.

Daily engagement rates (i.e., link sharing on Twitter/Reddit and search interest in Google Trends) is the instrument we use to measure these trends pre- and post-deplatforming. Figures 6 and 7 show the frequency and proportion of shares to these alternative platforms.

Model Results

Table 2 shows trends in engagement on BitChute, Gab, and Parler as compared to YouTube. Across these tables, we see mixed results. Engagement with Parler does seem to be significantly reduced following deplatforming–possibly a result of the removal of the platform from the Internet by its web-services provider. BitChute similarly sees less engagement in both social media sharing and search interest. Searches for BitChute do experience a significant increase in level following deplatforming – a level-increase that is two orders of magnitude higher than the suppressive trend. BitChute engagement remained above pre-event levels at the end of our sample period.

Interest in and engagement with Gab shows significantly different patterns: In all three spaces (Twitter, Reddit, and Google Trends), Gab sees a significant and immediate increase in engagement, suggesting much more interest in the platform soon after deplatforming in early January. Similar to BitChute, Gab does see a suppressive trend in the distance from treatment in both Twitter and Google Trends. As with search interest around BitChute, the increase in level of engagement is two orders of magnitude higher than the suppressive trend, again indicating several months would need to pass for interest to return to pre-treatment levels. More concerningly, Gab sees a significant trend *upward* in Reddit, suggesting interest in and engagement with Gab increased both immediately and continued to do so in the post-deplatforming period.

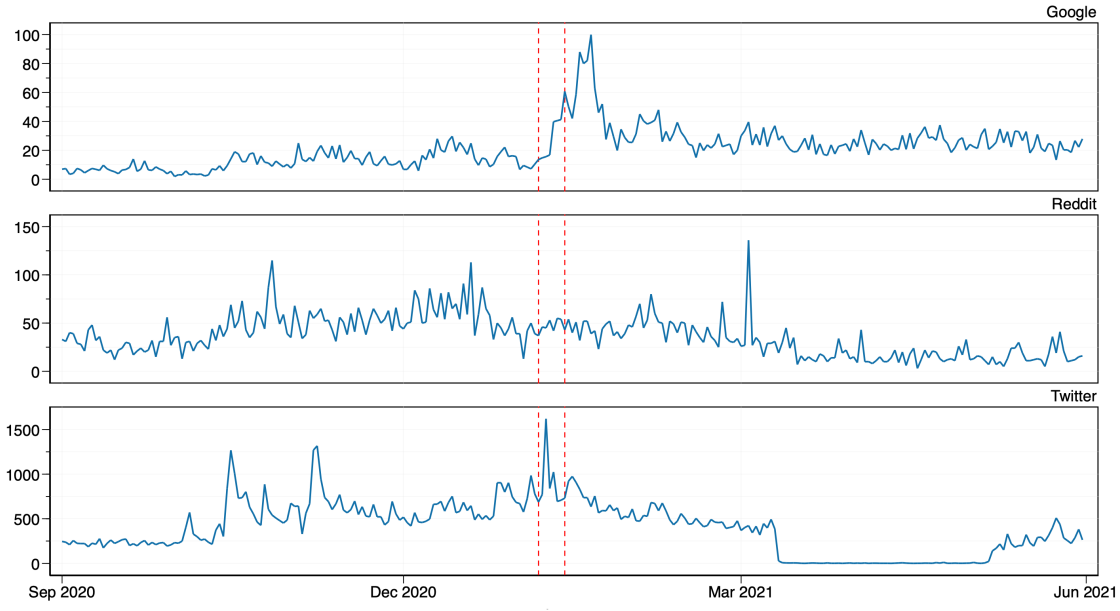
As a point of comparison, for YouTube we would expect to see limited changes in engagement after deplatforming. This expectation is consistent with findings in Twitter and Google Trends, where we see no significant changes over time. On Reddit, however, we do see a drop in the level of YouTube sharing and increase in trend, though these findings may be attributable to numerical issues in the model, as some factors present co-linearity

problems in the Reddit data. More analysis is needed here.

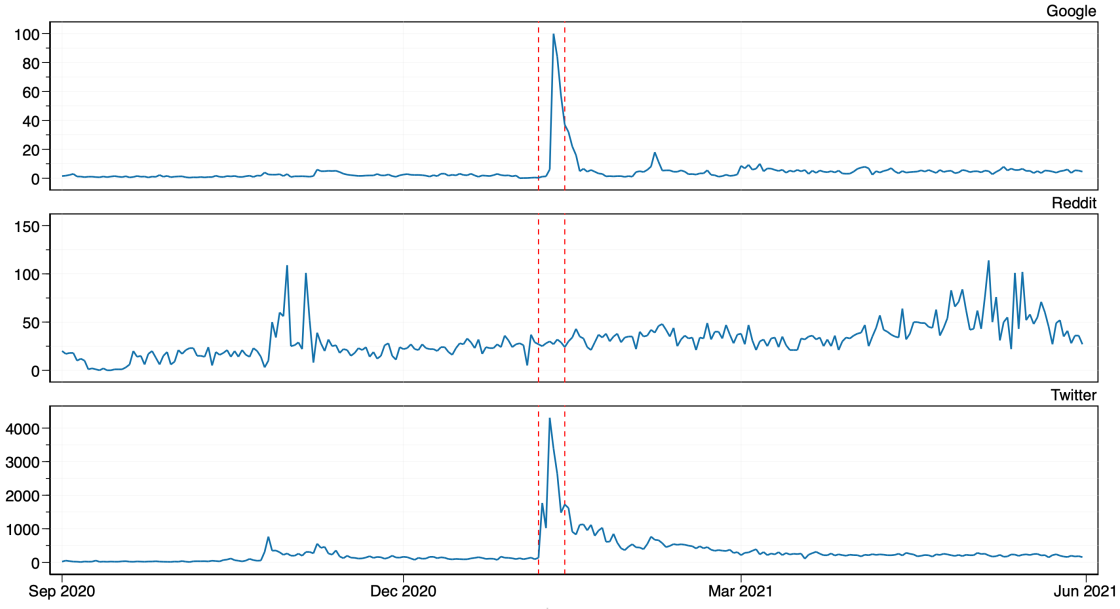
Table 2: Trends in links shared to alternative platforms and YouTube

		Twitter		Reddit		Google Trends	
BitChute							
Predictor		β	SE	β	SE	β	SE
Link-Sharing Vol.	l_t	0.0525 **	0.017	0.2259 ***	0.023	0.1897 ***	0.042
Lagged Sharing	$l_{BC,t-1}$	0.8674 ***	0.045	0.2264 **	0.056	0.7653 ***	0.050
News Coverage	$n_{BC,t}$	−0.0076	0.074	0.0922	0.079	0.0392	0.026
Rollout	$E(t)$	0.0275	0.100	−0.0605	0.055	0.1630 ***	0.042
Treatment	$T(t)$	0.1325	0.074	0.0547	0.085	0.1584 **	0.045
Dist. from Treat.	$d(t)$	−0.0098 **	0.003	−0.0112 ***	0.002	−0.0015 **	0.000
Observations			211		211		211
R^2			0.993		0.989		0.998
Gab							
Predictor		β	SE	β	SE	β	SE
Link-Sharing Vol.	l_t	0.0708 ***	0.017	0.1663 ***	0.023	0.0694 ***	0.013
Lagged Sharing	$l_{Gab,t-1}$	0.7640 ***	0.055	0.3003 **	0.098	0.6810 ***	0.054
News Coverage	$n_{Gab,t}$	−0.0180	0.044	−0.0281	0.042	−0.0056	0.027
Rollout	$E(t)$	0.9168 **	0.319	0.2128 *	0.095	0.9867 **	0.323
Treatment	$T(t)$	0.4183 **	0.124	0.2036 *	0.081	0.4507 ***	0.115
Dist. from Treat.	$d(t)$	−0.0033 **	0.001	0.0025 *	0.001	−0.0037 ***	0.001
Observations			211		211		211
R^2			0.996		0.990		0.975
Parler							
Predictor		β	SE	β	SE	β	SE
Link-Sharing Vol.	l_t	0.0427 **	0.016	0.0277 *	0.011	0.0362 *	0.018
Lagged Sharing	$l_{Parler,t-1}$	0.8831 ***	0.045	0.6922 ***	0.063	0.7792 ***	0.075
News Coverage	$n_{Parler,t}$	0.0300	0.054	0.1258	0.064	0.0429	0.052
Rollout	$E(t)$	−0.1127	0.594	−0.2299	0.580	0.6703 *	0.269
Treatment	$T(t)$	−0.5278 *	0.220	−0.6510 **	0.190	−0.0604	0.059
Dist. from Treat.	$d(t)$	0.0033	0.002	0.0039	0.002	−0.0016	0.001
Observations			211		211		211
R^2			0.990		0.851		0.918
YouTube							
Predictor		β	SE	β	SE	β	SE
Link-Sharing Vol.	l_t	0.5371 ***	0.050	0.5231 ***	0.011	0.0397	0.032
Lagged Sharing	$l_{YT,t-1}$	0.2792 ***	0.064	0.3358 ***	0.063	0.9491 ***	0.025
News Coverage	$n_{YT,t}$	0.0193	0.016	0.0186 ***	0.064	0.0130	0.010
Rollout	$E(t)$	−0.0631 **	0.020	−0.0200 **	0.580	−0.0120	0.013
Treatment	$T(t)$	−0.0421	0.022	−0.0195 ***	0.190	−0.0066	0.007
Dist. from Treat.	$d(t)$	0.0002	0.000	0.0002 ***	0.002	0.0000	0.000
Observations			211		211		211
R^2			1.000		1.000		1.000

$^*p < 0.05$;$^{**}p < 0.01$;$^{***}p < 0.001$



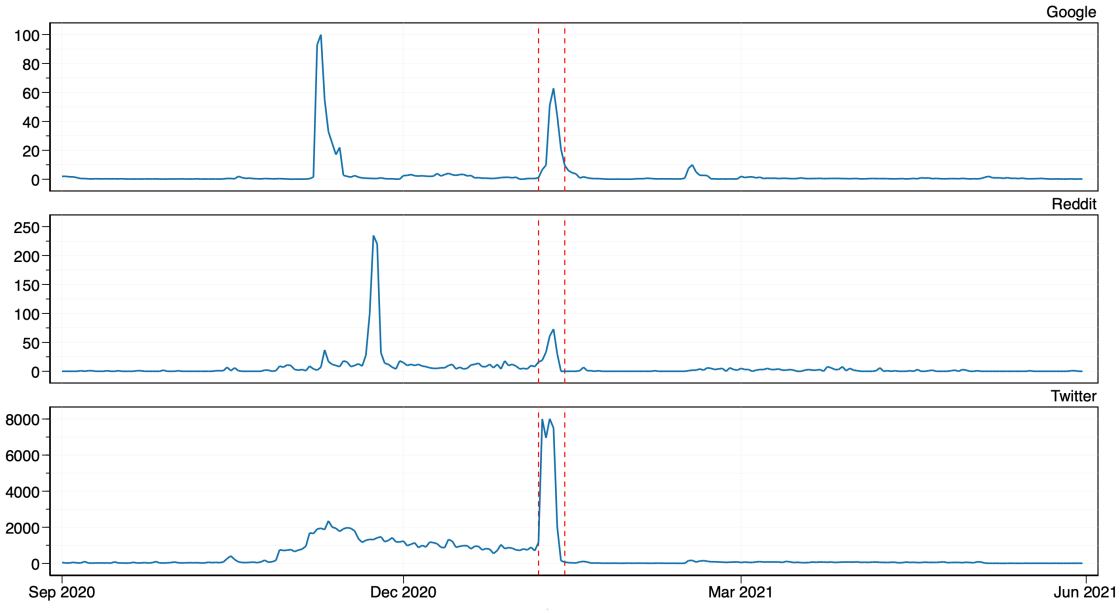
(a) BitChute Engagement



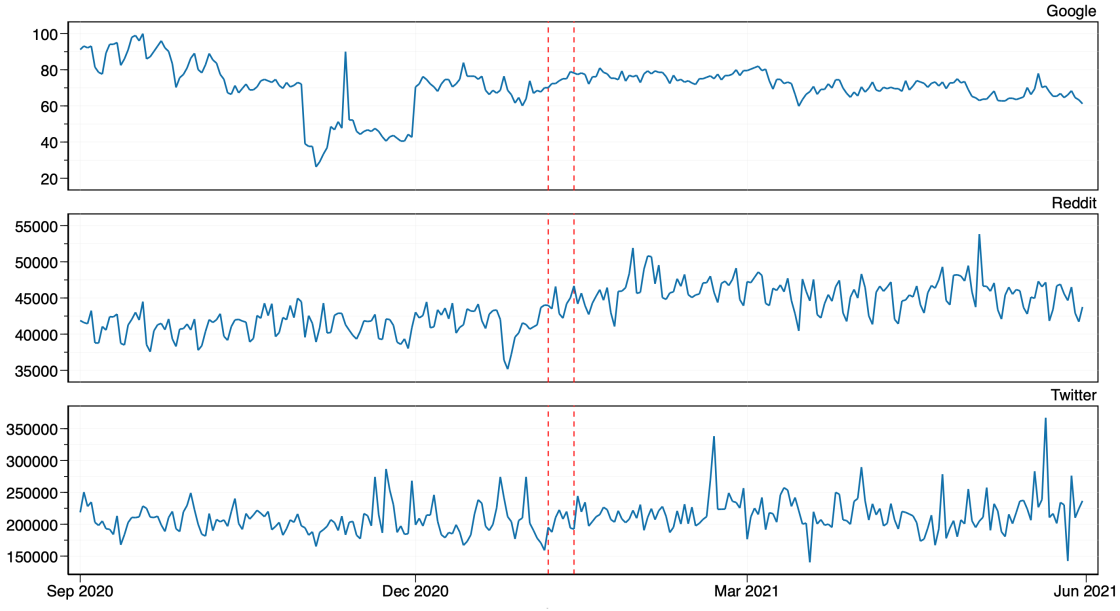
(b) Gab Engagement

Figure 6. Engagement with Alt-Tech Platforms and YouTube in Multiple Online Spaces.

Note. The red dashed lines illustrates the seven days between January 6 and when Twitter banned several thousand accounts, which includes when Facebook and Twitter banned President Trump. Data from Google, Twitter, and PushShift.io.



(a) Parler Engagement



(b) YouTube Engagement

Figure 7. Engagement with Alt-Tech Platforms and YouTube in Multiple Online Spaces.

Note. The red dashed lines illustrate the seven days between January 6 and when Twitter banned several thousand accounts, which includes when Facebook and Twitter banned President Trump. Data from Google, Twitter, and PushShift.io.

Content Shared on Various Platforms Before and After the Great Deplatforming

The analysis presented above provides insights about the temporal trends in platform usage and sharing. However, it is also important to understand the nature of conversation on platforms that gained users after January 6. In this section, we focus on Gab, the alt-social site that received the most attention by users on mainstream platforms. We focus on the content that was shared on Gab during this period, while paying particular attention to the prevalence of toxic content and hate speech. Note that because we are studying aggregate trends, we cannot distinguish changes in individual toxicity (i.e. people already on the platform behave differently) from compositional shifts (i.e. people who psoted toxic content elsewhere begin doing so on the new platform).

To get Gab data, we used the API of the Mastodon social network, on which Gab has been operating since July 2019. Our sample includes 17.5 million posts that were viewable on Gab's public timeline between September 2020 and April 2021.[4] Figure 8 shows overtime trends in user engagement on Gab in the three months before and after January 6. We use four measures of user engagement, including the daily number of posts, reblogs (which are similar to "retweets" on Twitter), replies, and favorites. We find that user engagement on Gab dramatically increased after the January 2021 deplatforming waves – a trend that lasted for about two months. The number of daily posts in our sample jumped from about 58 thousand daily posts in the month before the insurrection to about 129 thousand in the month after. By April 2021, the level of user engagement on Gab declined to about 74 thousand daily posts, but it stayed higher than it was before the attack on the Capitol.

To examine the nature of discourse on the platform, we measured the usage of keywords relating to several topics that were known to be popular among those who supported the storming of the U.S. Capitol. These include mentions of 'stop the steal,' posts describing the 2020 elections as fraud, and content talking about "big tech censorship."[5] Note that

[4]Mastodon is a decentralized, open source social network that allows smaller social media platforms to run their platforms on its servers. Gab's data is available through Mastodon's API. For more information see: https://docs. joinmastodon.org.

[5]We use the following keywords to measure mentioned of stop the steal: "stopthesteal", "stop the steal". To measure equating the 2020 elections with fraud, we use the following keywords: "voter-

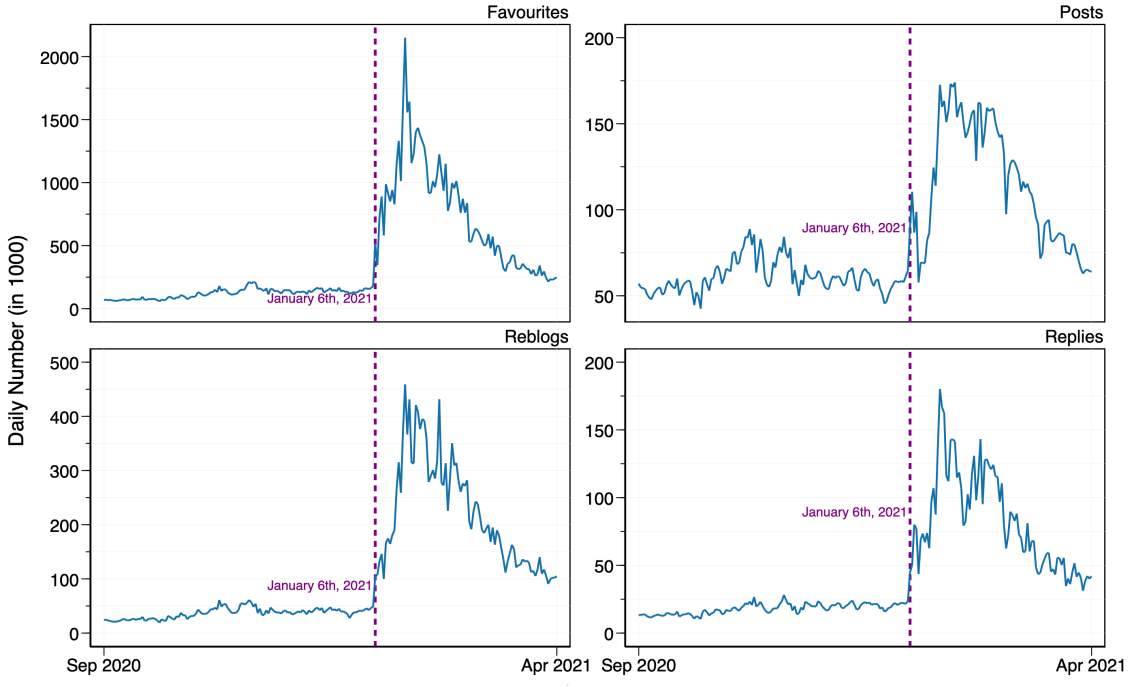


Figure 8. User Engagement on Gab

Note. The figure shows time series plots for daily user engagement on Gab (number of posts, re-blogs, replies, and favorites) in the three months before and after January 6. Data from Mastodon API.

this keyword based approach does not distinguish posts promoting a given term from those critiquing it.

We also use machine learning models to identify more general hateful and toxic rhetoric. The models draw on labeled data that consists of hateful posts targeting various minorities in the United States, as well as content promoting misogyny and endorsing white supremacy.[6] We created a hate speech index that summarizes these categories into one

fraud", "voter fraud", "electionfraud", "election fraud". To identify posts about big tech censorship, we use these terms: "censor*", "suspend*", "banned", "delet*", "de*platform*", "speech", "tech", "big tech*".

[6]The appendix provides more details on the models.

variable.

The top-right panel in Figure 9 shows trends in hate speech before and after the January 2021 events. The y-axis reflects the daily average of a hate speech index that measures hateful content targeting African Americans, Jews, Muslims, Asians, individuals from Latin American countries, immigrants, as well as content promoting misogyny and comments targeting the LGBTQ+ community. We find that hate speech on Gab significantly increased – and remained high – in the months following January 6th, 2021.[7]

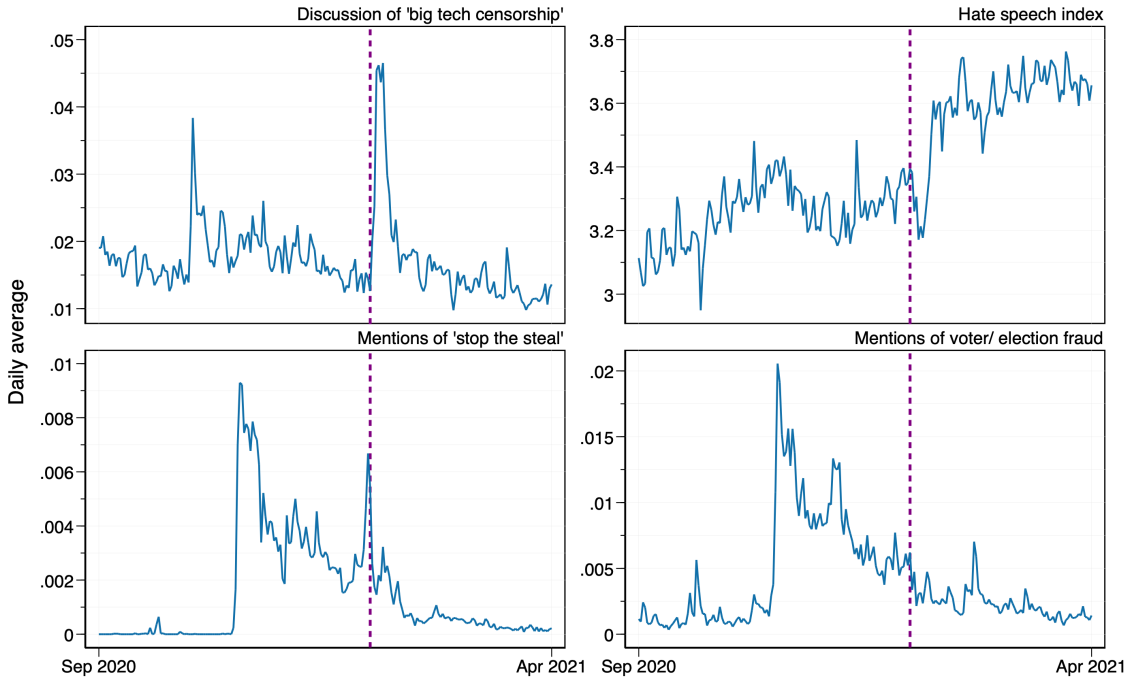


Figure 9. Discourse on Gab

Note. The figure shows time series plots for content expressing hate speech, as well as posts discussing censorship, voter fraud and 'stop the steal' that were posted on Gab in the three months before and after January 6. Data from Mastodon API.

[7]Appendix Figure A-3 shows similar patterns when using Google's Perspective API to measure hate speech and toxicity.

The other three panels in Figure 9 show overtime trends in content discussing voter fraud in the 2020 elections, mentions of 'stop the steal,' and discourse on 'big tech censorship' that tends to be popular in alternative social media platforms like Gab. We find that discourse on election fraud and content promoting the 'stop the steal' narrative increased dramatically in the aftermath of the US elections in November 2020, and declined over time. On the day of the Capitol riots, discourse on 'stop the steal' rose again, and then declined. As mass deplatforming events began to take place after January 6, content discussing censorship by big tech companies dramatically increased on Gab, a trend that lasted for several weeks. This finding is consistent with the central role discussions of censorship played in calls within Facebook to move to alternative platforms.

Did toxic content decrease on mainstream sites?

We also examine the prevalence of toxic content on mainstream platforms (from where actors were deplatformed). Using the same measure of hate speech that we used for Gab (see above), we examine how hateful language changed on Twitter and Reddit before and after the Great Deplatforming. Specifically, we estimate the change in use of hate speech terms using Equation 2:

$$y_t = \beta_1 Implementation + \beta_2 Post_1 + \beta_3 Post_2 + d_t + \epsilon_t, \tag{2}$$

Here, y_t is the percentage of posts on a given day which contain hate speech or white supremacist language, $Implementation$ is an indicator for days between January 6 and January 12 (the interim time period above), $Post_1$ is an indicator for the period January 13 to February 12, $Post_2$ is an indicator for the period February 13-March 12, and d_t are day of the week fixed effects to account for differential posting on specific days of the week. The baseline period is days from December 5 to January 5. The coefficients (β_1, β_2, and β_3) thus capture the change compared to the baseline period in the average percentage of posts that contain hate speech terms.

Figure 10 shows our findings for Reddit. We find that the percentage of posts containing any kind of hate speech spiked dramatically on Reddit in the week starting January

6, increasing by almost 10% from the baseline of 6.9% of posts as seen in Figure 10.[8] Hate speech on Reddit then dropped slightly below baseline in the first month post-deplatforming, before returning to baseline by the February-March period. There are slight increases in anti-Black and White Supremacist speech on Reddit during the week of January 6, but a return to baseline thereafter. Importantly, keyword based approach does not distinguish critiques of hate speech from promotion of it, so the usage spike on Reddit likely reflects a combination of pro- and anti-hate speech content.

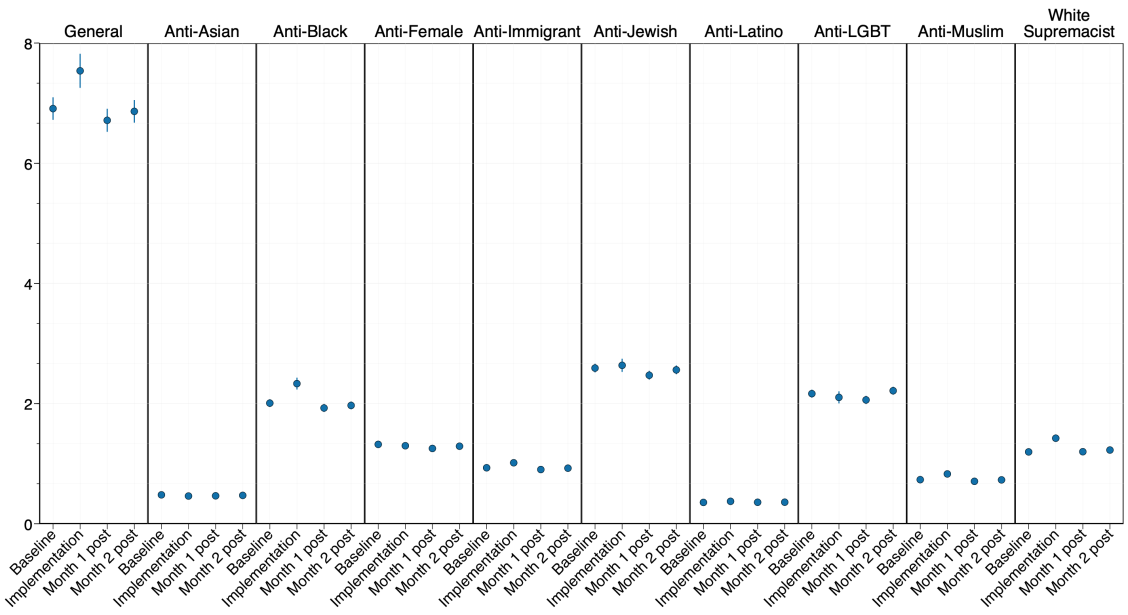


Figure 10. Great Deplatforming and different types of hate speech on Reddit

Note. The figure shows mean and 95% confidence interval on the percentage of posts containing different kinds of hate speech terms on Reddit around January 6 based on estimating equation 2. Data from PushShift.io. Hate speech terms from Mitts (2022).

Figure 11 shows that trajectory of hate speech on Twitter follows a different pattern than on Reddit.[9] We find that the percentage of tweets containing any kind of hate speech

[8]Full results in Appendix Table A-2.
[9]For full results see Appendix Table A-3.

spiked by more than 50% in the week starting January 6, from a baseline of 6.9% of posts. Aggregated hate speech discourse on Twitter remained elevated for the next two months, but this increase was not statistically significant. Several specific kinds of hate speech that spiked in the week of January 6 on Twitter did remain elevated through the next two months: Anti-Black hate speech increased approximately 15% from a baseline of 3%, while Anti-Female, Anti-LGBT and Anti-Muslim hate speech rose by a bit over 10%, from baselines of 2%, 2.2% and 1.2% respectively. As with Reddit, the spike in hate-speech related keywords likely reflects a combination of pro- and anti-hate speech content. Overall, though, hate-speech on Twitter did not decline notably after the post-January 6 deplatforming.

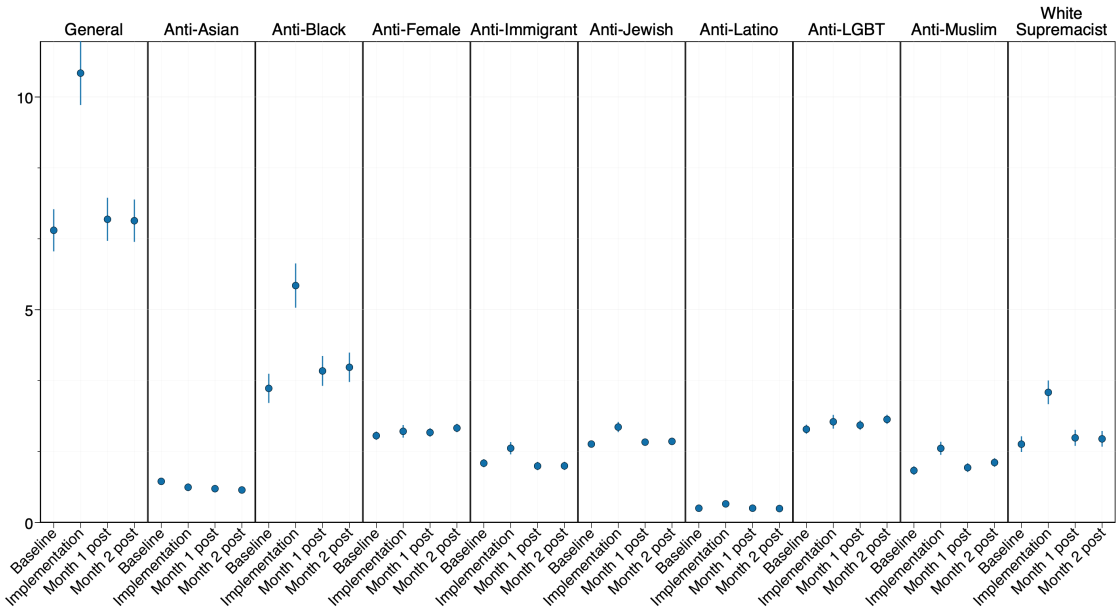


Figure 11. Great Deplatforming and different types of hate speech on Twitter

Note. The figure shows mean and 95% confidence interval on the percentage of posts containing different kinds of hate speech terms on Twitter around January 6 based on estimating equation 2. Data from Twitter decahose. Hate speech terms from Mitts (2022)

Limitations

Overall, we find that the 'Great Deplatforming' is not associated with a substantive increase in the quality of mainstream spaces (in some cases, as with Gab, we see the opposite). A key limitation of this analysis is that our metrics operate across large-scale collections of individuals rather than longitudinally, across a consistent set of accounts. Consequently, our analysis may miss individual-level changes in online behavior, especially if the target groups who are most likely to experience these interventions are small relative to the whole platform. That is, if the number of target accounts – e.g., those engaged in discussions of voter fraud or interacting with individuals who were banned – is small, changes in their behaviors may be dominated by the lack of change in population-level discussions. If that possibility is true, the absence of declines in toxic content or Gab-sharing on Twitter or Reddit might stem from coarse, insufficiently sensitive measures. It is unlikely that our results on Gab sharing or hate speech would not hold for those most closely engaged with deplatformed accounts or topics (they are the audiences for these deplatformed topics/accounts anyway). The limitation of working with samples of content, instead of following specific accounts, suggests a need for longitudinal analyses of specific populations, which we leave for future work.

Additionally, much of our observations herein are based on link-sharing within social media platforms, where a common hazard is counter-attitudinal sharing. That is, a well-known motivation for sharing content in political discourse is to denounce that content (Kim et al., 2020). As such, one might be concerned that the work herein is mainly counting users talking about how awful these alt-social platforms might be. This concern is especially common when studying "retweeting" on Twitter. While we note this possibility, such counter-attitudinal sharing is relatively rare (An et al., 2014), and for our purposes in studying increasing engagement with the alt-tech space, even if users' primary motivations are disparagement, such sharing does increase exposure to these spaces and facilitate easy linking into these alt-social platforms. Our analysis of search interest via Google Trends somewhat insulates us from this concern as well.

As we lay the foundation for this work, we have outlined contradictions in the literature around the efficacy of deplatforming. Some works demonstrate success (e.g.,

Chandrasekharan et al. (2017a), Jhaver et al. (2021), Buntain et al. (2021)), while others– this paper included–suggest caution (e.g., Rauchfleisch and Kaiser (2021), Mitts (2022)). Our descriptive analysis cannot speak to specific mechanisms, leaving open questions about how these results generalize to future instances of deplatforming and whether one might expect to see similar levels of backlash and engagement with alt-social spaces.

One way to think about the external validity of studies on content moderation is to consider deplatforming as part of a spectrum which ranges from subtle recommendation suppression as in Buntain et al. (2021), to harsher expulsion from a platform as in the Great Deplatforming, to full-on removal from the Internet (as with Parler). We study an example of the extreme side of such interventions. In that context, when deplatforming is sufficiently impactful as to garner substantial media coverage, the kinds of phenomena observed herein are more likely.

Importantly, large-scale instances of deplatforming are rare. De-recommendation, down-weighting, and "shadow banning" (Savolainen, 2022) appear to be much more common. More work is needed to understand these interventions, how they interact with media coverage, and under what conditions they might stimulate the kinds of backlash observed herein.

Conclusions

We study trends related to deplatforming across three dimensions: how users expecting to be deplatformed reacted, engagement on mainstream platforms and in search with alternative platforms, and how the discourse shifted on one alternative platform, Gab, and two major online spaces, Reddit and Twitter. We find that much of the movement to alternative platforms was announced on the mainstream platforms of Twitter and Facebook. Gab gained significant engagement across multiple spaces after January 6 compared to others. With the flood of new users, Gab became much more toxic, with hate speech rising to levels that were much higher than in previous months. On Twitter, hate speech spiked dramatically in the week of January 6th compared to the month before. And while overall hate speech on the platform returned to baseline levels, many specific categories remained elevated by 10-15% from the month of December, particularly anti-Black hate speech.

These results suggest that deplatforming can have complex effects. In prior work one of the authors argued that the deplatforming of English conspiracy theorist David Icke in early-May 2020 likely increased the prevalence and distribution of his material, and that the deplatforming in Fall 2020 had a short term impact that was recovered partially over-time (Innes and Innes, 2021). Our results are consistent with a displacement-diffusion dynamic in which deplatforming encourages a shift to the long-tail of niche platforms, leading to a negative shift in the tone of content on those platforms but no significant change in expression on the mainstream platforms.

Many aspects of American political discourse shifted after January 6. We provide evidence that the communities directly engaged in discussions supportive of that day's events responded to platform actions in strategic ways. They provided signposts to where the discussion could be continued, which coincided with a shift in the discourse on at least one alternative space. Understanding whether such dynamics are common, and thus the likely overall impact of deplatforming efforts, will require a combination of observing ecosystem level outcomes, as we begin to do here, with granular account-level analysis on multiple other events.

Acknowledgments

We thank Julia Ilhardt, Nilima Pisharody, Andrew Dawson and Hanjatiana Nirina Randrianarisoa for outstanding research support. We acknowledge support from the Carnegie Endowment for International Peace and Microsoft. Participants in the Carnegie Endowment - Princeton Symposium From Countering Behavior to Measuring Impact provided invaluable feedback. All errors are our own.

References

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, pages 187–195.

An, J., Quercia, D., and Crowcroft, J. (2014). Partisan sharing: Facebook evidence and societal consequences. *COSN 2014 - Proceedings of the 2014 ACM Conference on Online Social Networks*, (Figure 1):13–23.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. (2020). The Pushshift Reddit Dataset. In *arXiv preprint*.

Bernal, J. L., Cummins, S., and Gasparrini, A. (2018). The Use of Controls In Interrupted Time Series Studies of Public Health Interventions. *International Journal of Epidemiology*, 47(6):2082–2093.

Bernstein, M., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., and Vargas, G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

Bryanov, K., Vasina, D., Pankova, Y., and Pakholkov, V. (2022). The other side of deplatforming: Right-wing telegram in the wake of trump's twitter ouster. In Alexandrov, D. A., Boukhanovsky, A. V., Chugunov, A. V., Kabanov, Y., Koltsova, O., Musabirov, I., and Pashakhin, S., editors, *Digital Transformation and Global Society*, pages 417–428, Cham. Springer International Publishing.

Buntain, C., Bonneau, R., Nagler, J., and Tucker, J. A. (2021). Youtube recommendations and effects on sharing across online social platforms. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017a). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., and Gilbert, E. (2017b). You Can'T Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):31:1—-31:22.

Chang, J. P. and Danescu-Niculescu-Mizil, C. (2019). Trajectories of blocked community members: Redemption, recidivism and departure. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2:184–195.

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. (2014). Can cascades be predicted? pages 925–936. ACM.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial behavior in online discussion communities. *Proceedings of the 9th International Conference on Web and Social Media, ICWSM 2015*, pages 61–70.

Citron, D. (2021). It's time to kick trump off twitter. https://slate.com/technology/2021/01/twitter-kick-off-donald-trump.html. Accessed: October 27, 2021.

Coaston, J. (2018). Gab, the social media platform favored by the alleged pittsburgh shooter, explained. *Vox, October*, 29.

Cofnas, N. (2019). Deplatforming won't work. https://quillette.com/2019/07/08/deplatforming-wont-work/. Accessed: October 27, 2021.

Conger, K. (2021). Twitter, in widening crackdown, removes over 70,000 qanon accounts. *The New York Times*.

De Vynck, G. (2021). Youtube is banning prominent anti-vaccine activists. https://www.washingtonpost.com/technology/2021/09/29/youtube-ban-joseph-mercola/. Accessed: October 27, 2021.

Elegant, N. X. (2021). Tiktok banned trump before trump could ban tiktok. https://fortune.com/2021/01/11/tiktok-bans-trump-before-trump-bans-tiktok/. Accessed: October 27, 2021.

Fuchs, H. (2021). After capitol riot, elected officials under pressure back home. https://www.nytimes.com/2021/01/31/us/capitol-riot-local-politicians.html. Accessed: October 27, 2021.

Goforth, C. (2020). Banned by paypal and youtube, this alt-right comedian is back on paypal and youtube.

Innes, H. and Innes, M. (2021). Deplatforming disinformation: Conspiracy theories and their control. *Information, Communication, and Society*.

Jenkins, N. (2016). Twitter suspends conservative writer milo yiannopoulos.

Jhaver, S., Boylston, C., Yang, D., and Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30.

Kim, D. H., Jones-Jang, S. M., and Kenski, K. (2020). Why Do People Share Political Information on Social Media? *Digital Journalism*, 0(0):1–18.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press, 2 edition.

Mak, A. (2021a). Facebook is blocking trump from posting until at least inauguration day. https://slate.com/technology/2021/01/donald-trump-facebook-block-inauguration-day.html. Accessed: October 27, 2021.

Mak, A. (2021b). Twitter finally banned donald trump. https://slate.com/technology/2021/01/twitter-ban-donald-trump-byeeeee.html. Accessed: October 27, 2021.

Martineau, P. (2019). Facebook bans alex jones, other extremists—but not as planned. Accessed: October 27, 2021.

McIlroy-Young, R. and Anderson, A. (2019). From "welcome new gabbers" to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aaai conference on web and social media*, volume 13, pages 651–654.

Mencimer, S. (2016). Pizzagate shooter read alex jones. here are some other fans who perpetrated violent acts.

Mickle, T. (2021). Trump is still banned on youtube. now the clock is ticking. https://www.wsj.com/articles/trump-is-still-banned-on-youtube-that-could-change-11620293402. Accessed: October 27, 2021.

Mitts, T. (2022). Banned: How deplatforming extremists mobilizes hate in the dark corners of the internet.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *Proceedings of ICWSM*, pages 400–408.

Munn, L. (2021). More than a mob: Parler as preparatory media for the us capitol storming. *First Monday*.

Nouri, L., Lorenzo-Dus, N., and Watkin, A.-L. (2019). Following the whack-a-mole: Britain first's visual strategy from facebook to gab. *Global Research Network on Terrorism and Technology Paper*, (4).

Ohlheiser, A. (2016). Just how offensive did milo yiannopoulos have to be to get banned from twitter. *The Washington Post*.

Pater, J. A., Nadji, Y., Mynatt, E. D., and Bruckman, A. S. (2014). Just awful enough: the functional dysfunction of the something awful forums. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2407–2410.

Polantz, K., Atwood, K., Perez, E., and Rabinowitz, H. (2021). The january 6 attack on the u.s. capitol. Accessed: October 27, 2021.

Rauchfleisch, A. and Kaiser, J. (2021). Deplatforming the far-right: An analysis of youtube and bitchute. *Available at SSRN*.

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., and West, R. (2020). Does platform migration compromise content moderation? evidence from r/the_donald and r/incels. *arXiv preprint arXiv:2010.10397*.

Robertson, A. (2017). Neo-nazi site moves to dark web after godaddy and google bans. https://www.theverge.com/2017/8/15/16150668/daily-stormer-alt-right-dark-web-site-godaddy-google-ban. Accessed: October 27, 2021.

Rogers, R. (2020). Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229.

Rohlinger, D. A., Rose, K., Warren, S., and Shulman, S. (2023). Does the Musk Twitter Takeover Matter? Political Influencers, Their Arguments, and the Quality of Information They Share. *Socius: Sociological Research for a Dynamic World*, 9:237802312311521.

Romm, T. and Lerman, R. (2021). Amazon suspends parler, taking pro-trump site offline indefinitely. https://www.washingtonpost.com/technology/2021/01/09/amazon-parler-suspension/. Accessed: October 27, 2021.

Savolainen, L. (2022). The shadow banning controversy: perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6):1091–1109.

Seering, J., Kraut, R., and Dabbish, L. (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*, pages 111–125.

Stocking, G., Mitchell, A., Matsa, K. E., Widjaya, R., Jurkowitz, M., Ghosh, S., Smith, A., Naseer, S., and Aubin, C. S. (2022). The Role of Alternative Social Media in the News and Information Environment. Technical report, Pew Research Center.

Subramanian, C. (2021). A minute-by-minute timeline of trump's day as the capitol siege unfolded on jan. 6. https://www.usatoday.com/story/news/politics/2021/02/11/trump-impeachment-trial-timeline-trump-actions-during-capitol-riot/6720727002/. Accessed: October 27, 2021.

Sullivan, m. (2021). Facebook has deleted 19,500 groups tied to 'militarized social movements'. *Fast Company*.

Trujillo, M., Gruppi, M., Buntain, C., and Horne, B. D. (2020). What is bitchute? characterizing the ̈free speechälternative to youtube. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 139–140.

Twitter, Inc. (2021). Permanent suspension of @realdonaldtrump. Accessed: Nov 4, 2021.

Urman, A. and Katz, S. (2022). What they do in the shadows: examining the far-right networks on telegram. *Information, Communication & Society*, 25(7):904–923.

Zhou, Y., Dredze, M., Broniatowski, D. A., and Adler, W. D. (2019). Elites and foreign actors among the alt-right: The gab social media platform. *First Monday*.

Appendix

Measuring Hateful and Toxic Speech

We define hate speech by drawing on the Encyclopedia of Political Communication, whichdefines hate speech as:

> *"Comments containing speech aimed to terrorize, express prejudice and contempt toward, humiliate, degrade, abuse, threaten, ridicule, demean, and discriminate based on race, ethnicity, religion, sexual orientation, national origin, or gender... Also including pejoratives and group-based insults, that sometime comprise brief group epithets consisting of short, usually negative labels or lengthy narratives about an out group's alleged negative behavior."* (**?**)

Using Naive Bayes classifiers, we measured hate speech targeting African Americans, Jews, Muslims, Asians, individuals from Latin American countries, and immigrants, as well as content promoting misogyny and comments targeting the LGBTQ+ community. In addition to hate speech, we included in our measure endorsement of white supremacy. The models were trained on a labeled sample of 124,006 Twitter and Gab posts that were annotated by trained research assistants. We used 80% of the labeled data for training and evaluated performance on the remaining 20%. Since class imbalance was high, we re-balanced the training data by over-sampling posts from the minority category and under-sampling from the majority category.[10] Out-of-sample accuracy was 0.94, precision was 0.44, recall was 0.48, and the F1 score 0.46. The model was able to pick up hate speech, albeit with noise. This is partly because of the high class imbalance in our data. Measurement error would be the most problematic for our case if it was trending over time. However, as Figure A-1 shows, there is little variation in the error rate over time. To create the figure, we used our test data to measure, for each date, the average rate of correct classifications. The average error rate of 0.06 stays largely similar throughout the period of the test data.

[10] We used the ovun.sample() function from the ROSE package in R, setting the method of over and under-sampling to be "both." This increased the positive class by 37,047 observations and decreased the negative class by 51,621 observations. The re-balanced training set included 42,321 over-sampled positive cases and 42,309 under-sampled negative ones.

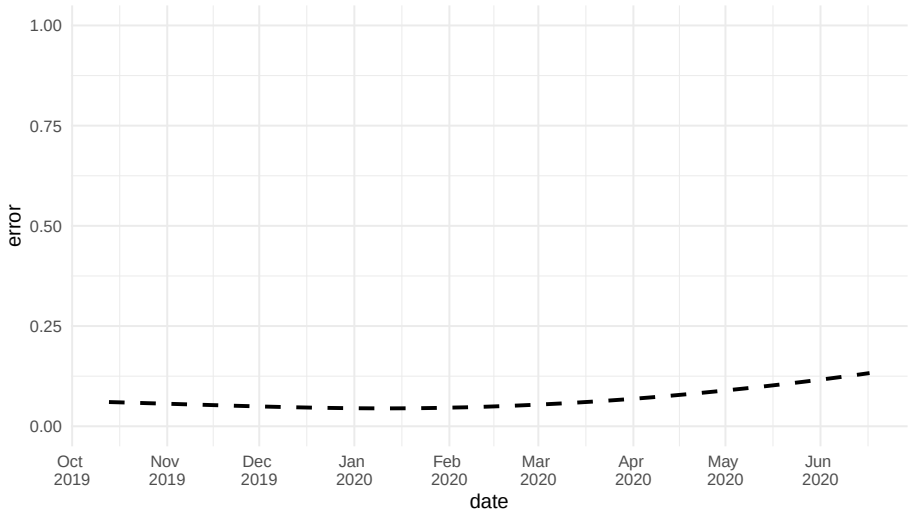


Figure A-1. Average Error Over Time

Thus, the "true" level of hate speech in the predicted class seems to be stable over time.

Table A-1 shows the words that were identified by the models to be predictive of each hate speech category. Examining these words is useful for understanding the themes that were captured in each topic. As expected, much of the predictive content includes words that refer to the targeted groups. But a close look shows particular themes that characterize hate speech for each category. For example, in anti-Black hate speech, racist slurs are often used, as well as references to the Black Lives Matter movement. Hateful posts against Muslims include references to terrorism and violence. Anti-Asian content is strongly linked to discussion on the COVID-19 pandemic. And posts expressing hate towards Latino communities tend to refer to illegal immigration. Interestingly, content endorsing white supremacy is strongly predicted with hashtags relating to the QAnon conspiracy theory, such as #wwg1wga„ #q, and #thegreatawakening. This is likely driven by the timing of the data collection, which took place during a peak in the popularity of the QAnon movement.

Since our Naive Bayes model predicted hate speech with a high degree of noise, we replicated our results with a more sophisticated deep learning model with a long-short

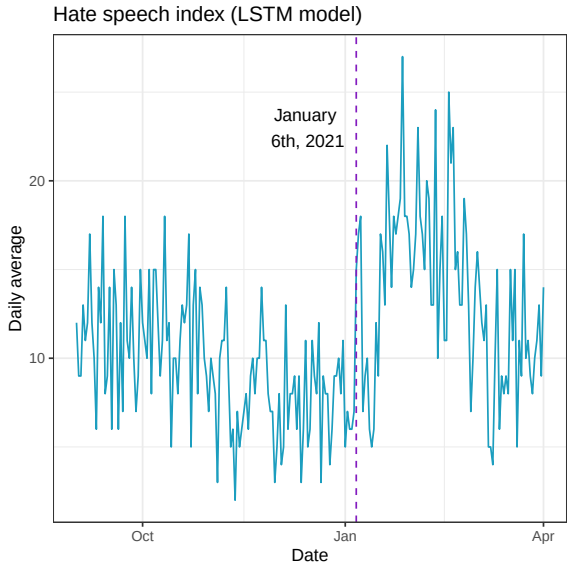


Figure A-2. Hate Speech Index (LSTM Model)

Note. The figure presents an alternative measure of our hate speech index, using an LSTM deep learning model to predict hate speech.

term memory (LSTM) layer.[11] As before, we trained the model on 80% of the data and evaluated it on the remaining 20%. The LSTM model performed much better than the Naive Bayes model. The overall accuracy was 0.88, recall 0.87, precision 0.88 and F1 0.88. Since the model required significantly more computational power, we used a random sample of 100,000 posts from our data to examine whether patterns of hate speech look similar to our baseline, Naive Bayes model. Figure A-2 shows very similar time trends to our findings in Figure 9, increasing our confidence that our Naive Bayes model, while noisy, is able to capture trends in toxic content over time.

[11] We used the Keras package in R for this purpose. The model is built on a neural network with an embedding layer, an LSTM layer, and an output dense layer that is used for prediction.

Table A-1: Most Predictive Words in Each Category

Category	*Terms*
Anti-Black	black, white, people, n***, blacks, police, whites, matter, lives, racist, america, f***, sh***
Anti-Jewish	jew, jewish, people, white, a.d, world, israel, media, war, america, god, holocaust, evil
Anti-Muslim	muslim, islam, people, islamic, death, country, police, government, terrorist, kill, democrats
Anti-Latinx	mexico, america, border, cartel, illegal, jorge, back, ramos, immigration, gangs
Anti-Asian	china, virus, coronavirus, chinese, communist, health, wuhan, pandemic, covid-19, flu
Anti-immigrant	illegal, people, immigration, country, muslim, illegals, immigrants, america, migrants, eu, back
Misogyny	women, men, white, woman, bitch, love, children, #maga, child, sh***, good, young, f***
Anti-LGBTQ+	people, women, men, gay, sex, children, transgender, court, sexual, gender, trans, fag***, god, liberal
Pro-white supremacy	#maga, #greatawakening, #wwg1wga, america, #q, #kag, whites, #trump2020, #thegreatawakening, #qanon, trump, slave, #trusttheplan, #redpill

Deplatforming and hate speech on Reddit & Twitter

This section contains additional results referenced in the main text.

Table A-2: Great Deplatforming and different types of hate speech on Reddit

VARIABLES	(1) General	(2) Anti-Asian	(3) Anti-Black	(4) Anti-Female	(5) Anti-Immigrant	(6) Anti-Jewish	(7) Anti-Latino	(8) Anti-LGBT	(9) Anti-Muslim	(10) guns	(11) White Supremacist
Implementation	0.628***	-0.0197	0.324***	-0.0228	0.0833***	0.0455	0.0179***	-0.0617	0.0953***	0.351***	0.228***
	(0.135)	(0.0124)	(0.0469)	(0.0201)	(0.0204)	(0.0522)	(0.00649)	(0.0477)	(0.0186)	(0.0319)	(0.0294)
Month 1 post	-0.195**	-0.0163**	-0.0809***	-0.0681***	-0.0296**	-0.119***	0.00243	-0.104***	-0.0277**	-0.0175	0.00315
	(0.0811)	(0.00744)	(0.0282)	(0.0121)	(0.0123)	(0.0314)	(0.00390)	(0.0287)	(0.0111)	(0.0191)	(0.0177)
Month 2 post	-0.0462	-0.00898	-0.0387	-0.0311**	-0.00683	-0.0295	0.00369	0.0486	-0.00370	-0.0144	0.0313*
	(0.0830)	(0.00762)	(0.0288)	(0.0124)	(0.0126)	(0.0321)	(0.00399)	(0.0293)	(0.0114)	(0.0196)	(0.0181)
Constant	6.910***	0.481***	2.008***	1.321***	0.932***	2.591***	0.355***	2.165***	0.734***	0.841***	1.196***
	(0.0956)	(0.00877)	(0.0332)	(0.0142)	(0.0145)	(0.0370)	(0.00460)	(0.0338)	(0.0131)	(0.0226)	(0.0208)
N	100	100	100	100	100	100	100	100	100	100	100
R^2	0.328	0.162	0.467	0.305	0.281	0.213	0.103	0.278	0.345	0.629	0.441
Mean of DV	6.926	0.487	1.985	1.277	0.934	2.550	0.353	2.136	0.733	0.876	1.234
SD of DV	0.376	0.0309	0.147	0.0551	0.0551	0.134	0.0157	0.128	0.0524	0.119	0.0898

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A-3: Great Deplatforming and different types of hate speech on Twitter

VARIABLES	(1) General	(2) Anti-Asian	(3) Anti-Black	(4) Anti-Female	(5) Anti-Immigrant	(6) Anti-Jewish	(7) Anti-Latino	(8) Anti-LGBT	(9) Anti-Muslim	(10) guns	(11) White Supremacist
Implementation	3.693***	-0.139***	2.415***	0.102	0.352***	0.399***	0.102***	0.176**	0.521***	2.293***	1.218***
	(0.357)	(0.0432)	(0.247)	(0.0705)	(0.0686)	(0.0571)	(0.0222)	(0.0771)	(0.0726)	(0.223)	(0.133)
Month 1 post	0.258	-0.172***	0.409***	0.0761*	-0.0662	0.0448	0.000835	0.0956**	0.0679	0.226*	0.148*
	(0.214)	(0.0259)	(0.149)	(0.0424)	(0.0412)	(0.0343)	(0.0133)	(0.0463)	(0.0436)	(0.134)	(0.0797)
Month 2 post	0.225	-0.202***	0.495***	0.180***	-0.0627	0.0632*	-0.00768	0.232***	0.188***	0.538***	0.125
	(0.219)	(0.0265)	(0.152)	(0.0434)	(0.0421)	(0.0351)	(0.0137)	(0.0474)	(0.0446)	(0.137)	(0.0815)
Constant	6.868***	0.968***	3.153***	2.041***	1.395***	1.845***	0.337***	2.192***	1.223***	1.225***	1.842***
	(0.253)	(0.0306)	(0.175)	(0.0499)	(0.0485)	(0.0404)	(0.0157)	(0.0546)	(0.0514)	(0.158)	(0.0938)
N	100	100	100	100	100	100	100	100	100	100	100
R^2	0.570	0.463	0.528	0.189	0.352	0.372	0.243	0.236	0.411	0.566	0.506
Mean of DV	7.392	0.874	3.654	2.078	1.431	1.884	0.343	2.328	1.310	1.649	1.999
SD of DV	1.243	0.135	0.822	0.179	0.194	0.165	0.0583	0.201	0.216	0.775	0.431

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A-4: Summary Data from Crowdtangle for top 30 accounts with Rumble links

Page/Group Name	Posts	Political posts	Total	Affiliation	Type
Rumble	4080	1	4081	n/a	Page
Dinesh D'Souza	695	597	1292	Conservative	Page
Dan Bongino	629	595	1224	Conservative	Page
We The 74,000,000	473	330	803	Conservative	Group
Conservative News Network	507	262	769	Conservative	Page
FOX NEWS with Tucker Carlson	485	250	735	Conservative	Group
La Caja de Pandora	584	49	633	n/a	Group
L'Eretico	476	133	609	Conservative	Group
Trump Team Sweden	349	249	598	Conservative	Group
It's A Pittie & Bully Thing	591		591	n/a	Page
It's a Pug Thing	563		563	n/a	Page
Støttegruppe for president Donald J. Trump. USA	322	238	560	Conservative	Group
Rumble Dogs	524		524	n/a	Page
Believe It or Not You Decide	312	144	456	Conservative	Group
Bongino Report	230	224	454	Conservative	Page
The World Evangelistic Union	442		442	n/a	Group
JESUS IS A CONSUMING FIRE	430		430	n/a	Group
GOSPEL HOUSE	423		423	n/a	Group
Rumble.com friends	415	3	418	n/a	Group
Rumble Babies & Kids	402		402	n/a	Page
Glorious And Free	372	30	402	Conservative	Group
Arizona Legislative District 2 Republican Party	234	150	384	Conservative	Page
La Caja de Pandora | VIDEO	353	29	382	n/a	Page
Rumble Cats	380		380	n/a	Page
The Maine Conservative Voice	186	179	365	Conservative	Page
Trump Keep America Great 2020	125	238	363	Conservative	Group
Trump 2024	186	150	336	Conservative	Group
Flawed Ink	213	123	336	Conservative	Page
JESSE WATTERS FAN PAGE	198	134	332	Conservative	Group
Christina Aguayo News	211	121	332	Conservative	Page

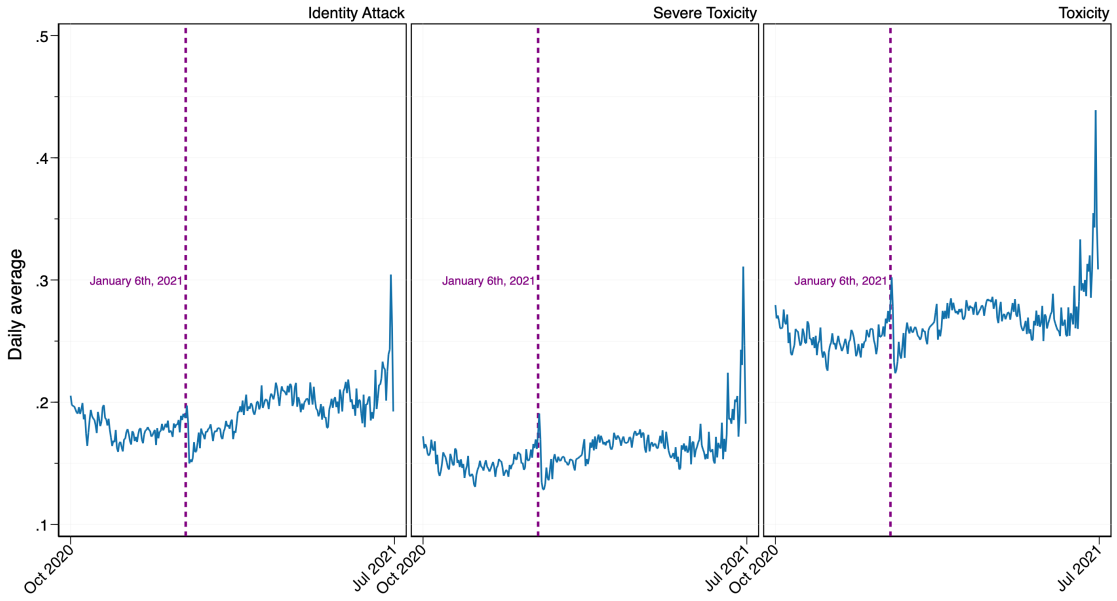


Figure A-3. Discourse on Gab (Alternative Measures of Hate Speech)

Note. The figure shows time series plots for content expressing hate speech, measured with Google's Perspective API, that were posted on Gab in the three months before and after January 6. *Identity attack* includes "Negative or hateful comments targeting someone because of their identity." *Toxicity* consists of "rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion." *Severe toxicity* is defined as a "very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective. This attribute is much less sensitive to more mild forms of toxicity, such as comments that include positive uses of curse words." Source: https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages