Banned: How Deplatforming Extremists Mobilizes Hate in the Dark Corners of the Internet

Tamar Mitts*

October 18, 2021

This work is part of my book project:
Digital Counterterrorism: Why Combatting Online Extremism is So Hard – And What Can Be Done About It

Abstract

In recent years, the world has seen a rapid increase in the use of social media platforms by violent extremists. Hate groups espousing radical ideologies have been using online platforms to communicate, disseminate propaganda, and in some cases, plan violent acts. In response, social media companies have upped their efforts to take down content and prevent the spread of hate speech on their platforms. While these actions reduced the availability of extremist content on mainstream social media, little is known about what happens to suspended individuals after being deplatformed. This project sheds light on the effects of deplatforming among online communities affiliated with the far-right in the United States. Analyzing cross-platform data that includes information on individuals who have accounts both on Twitter (a mainstream platform) and Gab (a fringe platform favored by far-right extremists), I find that Twitter suspensions increase engagement with hate speech on Gab. I discuss several approaches that can help mitigate radicalization on fringe platforms.

*Assistant Professor, School of International and Public Affairs, Columbia University. Email: tm2630@columbia.edu.

1 Introduction

> *"We suck at dealing with abuse and trolls on the platform and we've sucked at it for years... We're going to start kicking these people off right and left and making sure that when they issue their ridiculous attacks, nobody hears them."* – Twitter CEO Dick Costolo, February 2015

> *"Free speech includes speech that is critical, mocks, or confronts taboo topics that will offend someone somewhere in regards to race, politics, history, science, theology, philosophy, and any other topic of discussion. Deal with it."*– Gab CEO Andrew Torba, April 2021

How should we combat hate speech and extremism on social media platforms? Over the past decade, violent extremist groups have dramatically increased their use of online platforms for propaganda dissemination and recruitment. Groups like the Proud Boys and other far-right militias have been able to attract sympathizers through elaborate online campaigns, inspiring an alarming growth in white supremacist violence. Attacks such as the Pittsburgh synagogue shooting, the Christchurch mosque attack in New Zealand, and the recent storming of the U.S. Capitol have all been linked to extremist online content, leading to a strong public push to regulate social media platforms (Beach, 2019; Thompson, 2019; Romm, 2021).

While initially reluctant to take an active role in content moderation, since the mid-2010s many social media companies began deleting posts and banning accounts promoting hate speech and violence. Facebook, for example, expanded its content moderation policies and built algorithms to block extremist content before it is uploaded. Google designed tools to automatically detect and take down content promoting extremism, and Twitter stepped up its efforts to suspend accounts disseminating propaganda and hate speech. Figure 1 displays data from transparency reports by these companies, showing that some of the largest increases in takedowns since 2019 have been related to hate speech.

Yet despite the increase in content moderation, far-right extremism and hate-based violence have continued to rise, particularly in the United States. What can explain the growth in internet-inspired extremism, in spite of recent crackdowns on extremist content on mainstream social me-

dia? I suggest that part of the explanation may lie in the growing availability of alternative platforms that do not moderate content promoting hate. Gab, for example, is a social media site that takes pride in its openness to all types of content, including hate speech. Parler is another platform that, by design, takes very little action against extremist and hateful posts. These new platforms have become increasingly popular among far-right communities who are banned from mainstream online platforms. By January 2021, web traffic to these platforms has increased by over 600%.[1]

Since current research on content moderation focuses almost exclusively on mainstream social media (Chandrasekharan et al., 2017; Thomas and Wahedi, 2021), we have very little knowledge on how actions to moderate content shape engagement with extremist content on smaller, alternative platforms. On one hand, shutting down accounts may be effective at quelling online extremism across the internet as a whole, as those who promote extremism and hate might find it hard to attract an audience on smaller online forums. On the other hand, post-deplatforming migration to alternative platforms could inspire deeper engagement with extremist content and hate speech. In such cases, deplatforming does not solve the problem of online extremism, it just displaces it elsewhere.

In this paper, I examine what happens on the fringes of the internet when mainstream social media platforms engage in content moderation. Drawing on unique cross-platform data that includes information on individuals who have accounts on Twitter and on Gab, I study how Twitter suspensions shape engagement with hate speech on Gab. To gather data on deplatforming, I developed a real-time tracker that identified instances of user suspensions from Twitter in a sample of about 30 thousands individuals who have accounts on both Twitter and Gab. To measure hate speech, I trained several machine learning models to identify posts conveying hate towards various identity groups in the United States, as well as content endorsing white supremacy and far-right conspiracy theories.

Since the exact timing of Twitter suspensions is arbitrary and not systematically linked to the

[1]Web traffic to Gab steadily decreased since January 2021, but remained higher than average. `https://www.similarweb.com`

content that users post on the platform, I am able to treat users' deplatforming events as plausibly-exogenous shocks. Estimating several difference-in-differences models that compare overtime trends in hate speech posted on Gab by users who were banned from Twitter to those that are not, I find that deplatforming is associated with a significant increase in hate speech on Gab. A large portion of the increase relates to content promoting white supremacy and hate towards minority groups in the United states.

These findings highlight the importance of evaluating content moderation policies from the perspective of the internet ecosystem as a whole. While deplatforming extremists from mainstream social media reduces their presence on the banning platforms, it can, at the same time, mobilize hate on other platforms. Since mobilization on fringe platforms has already shown inspire offline harm—as has been the case with the storming of the U.S. capitol, for example—it is imperative to not ignore online activity on smaller, less moderated platforms.

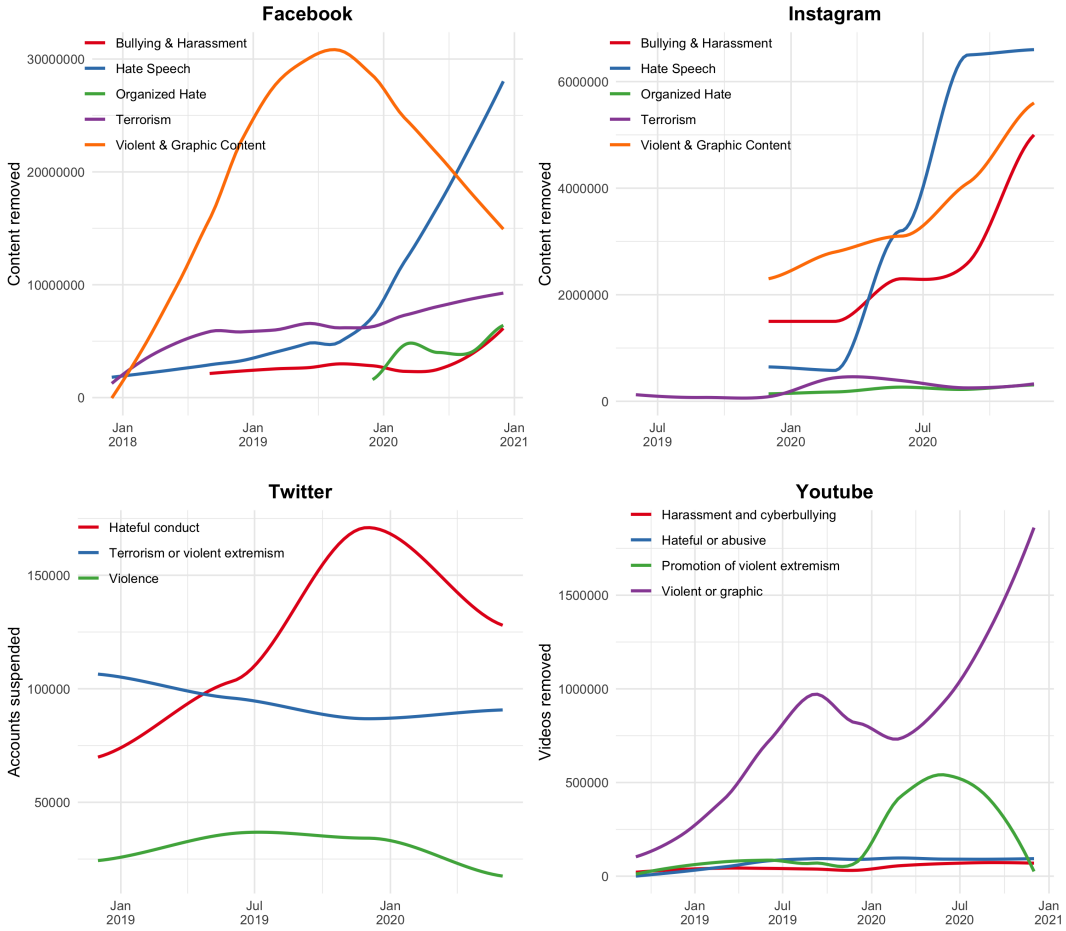
2 Recent Trends in Content Moderation

Mainstream social media platforms have engaged in the takedown of extremist content for several years. The first major effort began around 2015, and focused on banning content linked to extremist groups like the Islamic State and Al-Qaeda. More recently, moderation efforts have expanded to content endorsing white supremacy and hate speech, where platforms have taken action against thousands of accounts promoting far-right extremism and violent conspiracy theories.

Public information on enforcement actions became available around 2018, when several social media platforms began publishing transparency reports. The reports provide summaries of the amount of content that was taken down in different time periods, including, in some cases, the number of accounts that have been deplatformed. Figure 1 shows these data for Facebook, Instagram, Twitter, and YouTube. The colored lines represent different content moderation categories, which have somewhat different definition across platforms. Since 2019, Facebook, Instagram, and Twitter have all dramatically increased the takedown of content related to hate speech, harassment

and violence. While YouTube has upped its efforts to take down violent videos, its moderation of extremist content has declined in the latter half of 2020 and it has taken relatively little action against hateful or harassing content. After the storming of the U.S. capitol in January 2021, these platforms reported that they have taken additional actions against far-right content, which include the takedown of over 20,000 Facebook groups and pages and over 70,000 accounts on Twitter (Conger, 2021; Sullivan, 2021).

Figure 1: Content Moderation by Mainstream Platforms



Note: The figure presents data from transparency reports on content moderation by Facebook, Instagram, Twitter and YouTube.

3 The Debate Over Deplatforming

Even though enforcement actions have been increasing over the years, there is still an ongoing debate over deplatforming and its effectiveness. The debate often revolves around how one defines 'success' and whether the definition consists of within-platform effects or cross-platform migration and spillovers.

Deplatforming works

The pro-deplatforming camp tends to emphasize the importance of keeping mainstream social media 'clean' of extremist activity. One of its main arguments is that deplatforming takes away the stage that extremists use to publicize their cause and therefore decreases public exposure to extremist messaging. A large literature on terrorism has emphasized the importance of publicity for militant groups, both for making their cause known, as well as for attracting potential recruits (Kydd and Walter, 2006; Enders and Sandler, 2011). By blocking access to mainstream social media, deplatforming disrupts extremists' ability to disseminate propaganda and engage individuals who might have never heard about them outside of the online world.

Deplatforming can also negatively impact militant groups' operations. In recent years, extremists have been using social media platforms not only for content dissemination, but also for coordination and organization of violent activity. Blocking extremists from online platforms can therefore inhibit their ability to coordinate. A recent study conducted on Reddit illustrates this dynamic. In 2015, Reddit enacted a new anti-harassment policy that led to the banning of several groups promoting hate speech on the platform. Before the ban, these groups were active at promoting hate speech and attracting crowds to their toxic circles. After the ban, most members of these groups stopped using Reddit, and those who stayed on the platform dramatically reduced their engagement with hate speech (Chandrasekharan et al., 2017). From a within-platform perspective, these takedowns were successful, as they reduced the level of hate speech on Reddit.

Finally, banning extremists from mainstream social media can also dramatically shrink their audience on alternative platforms. Even though smaller social media platforms often welcome

extremists who are banned from the large platforms, they are not able to provide the same level of user engagement. As a result, extremist groups that are deplatformed from mainstream social media often experience a dramatic drop in followers, even if they recover their online presence on alternative platforms. This is what happened to Britain First, a U.K.-based hate group that was banned from Facebook in 2018. Before it was deplatformed, the group had over 1.8 million followers on Facebook who actively shared the group's content on the platform. After its suspension, the group migrated to Gab and Telegram, but was not able to attract the same amount of followers (Nouri, Lorenzo-Dus and Watkin, 2019).

Deplatforming does not work

On the other side of the debate are those who criticize current deplatforming efforts. The core of the disagreement revolves around the definition of success. While the pro-deplatforming camp tends to focus on mainstream social media and on within-platform effects—e.g., decreases in extremist activity on the banning platforms—the second camp advocates evaluating deplatforming from the cross-platform perspective. Critics of deplatforming argue that even though it reduces extremist activity on mainstream social media, hate groups can still successfully coordinate on alternative platforms (Cofnas, 2019; Bennett, 2018). Thus, focusing only on the large, mainstream platforms ignores important dynamics taking place on alternative online sites, which can have real-world consequences.

One example is the coordination of far-right activists on fringe platforms in the wake of the January 6 storming of the U.S. Capitol. While mainstream platforms like Facebook and Twitter banned many of their accounts in the months preceding the insurrection, activists were nonetheless able to coordinate on platforms such as Gab, 4chan, and Parler. In the weeks preceding the insurrection, these actors strongly promoted the event on alternative platforms, and some even live-posted their actions during the riot (Frenkel, 2021; Heilweil and Ghaffary, 2021). Since systematically monitoring fringe platforms is often difficult, some worry that pushing extremists to darker corners of the internet would make it harder for authorities to track them in order to prevent

offline harm (Greer, 2020).

In addition, in some cases, the actual act of deplatofrming can motivate users to further engage with the banned content. This can stem from feeling rejected and seeking comfort among communities of like-minded people, or from a sense of injustice that content bans are biased against certain groups. In the U.S. context, some have argued that Big Tech has a "pro-liberal bias," which leads companies to deplatform users and groups associated with the political right, but not those on the left side of the spectrum (Smith, 2018). While companies dispute these claims, many social media users strongly believe this is happening.[2] The mobilizing effect of deplatforming was also observed by Roberts (2018), who studied how social media users in China reacted to censorship. Analyzing the online behavior of users who whose content was censored, she found that they were more likely to mobilize against the government and seek the censored information than those who were not censored. While the contexts are very different, the mechanisms driving the behavior of Chinese users after online censorship might be similar to those that motivate action by banned users in the United States.

Finally, a broader argument made by those who oppose deplatforming is that large-scale bans can disrupt healthy political discourse in democratic societies. Social media platforms, at least in theory, can facilitate a productive debates in the marketplace of ideas, where those across the political spectrum exchange various perspectives on a range of policy issues (Schroeder, 2018). While current trends in online political discourse are far from this ideal,[3] some worry that deplatforming will eventually lead to a world with many different siloed communities that 'live' in separate online platforms and have little interaction with each other.

As of now, the debate is unresolved, partly because current research focuses almost exclusively on within-platform effects. My goal in this study is advance our knowledge on the cross-platform consequences of deplatforming, by examining how banning users from one social media platform affects their activity on another. Focusing on far-right extremism in the United States, I study how

[2]For some examples from the Gab platform see Appendix Table A3.

[3]See Persily and Tucker (2020) for a recent review.

suspending users from Twitter shapes online activity on Gab. The next section explains how I collected cross-platform data on these two platforms.

4 Data and Measurement

To understand the cross-platform effects of deplatforming, we need: (i) a plausibly exogenous measure of user suspensions from social media platforms, (ii) data on user behavior on alternative platforms, and (iii) a way to observe engagement with hateful, extremist content. Since social media platforms do not systematically release such information, I developed a prospective data collection infrastructure, which has been tracking, since October 2019, user behavior on two online platforms: Twitter and Gab. Prospective collection allows capturing instances of user suspension in real time, and provides large amounts of textual data that can be used to measure engagement with hate speech across platforms. Figure 2 provides an overview of the data collection, which I explain in more detail below.

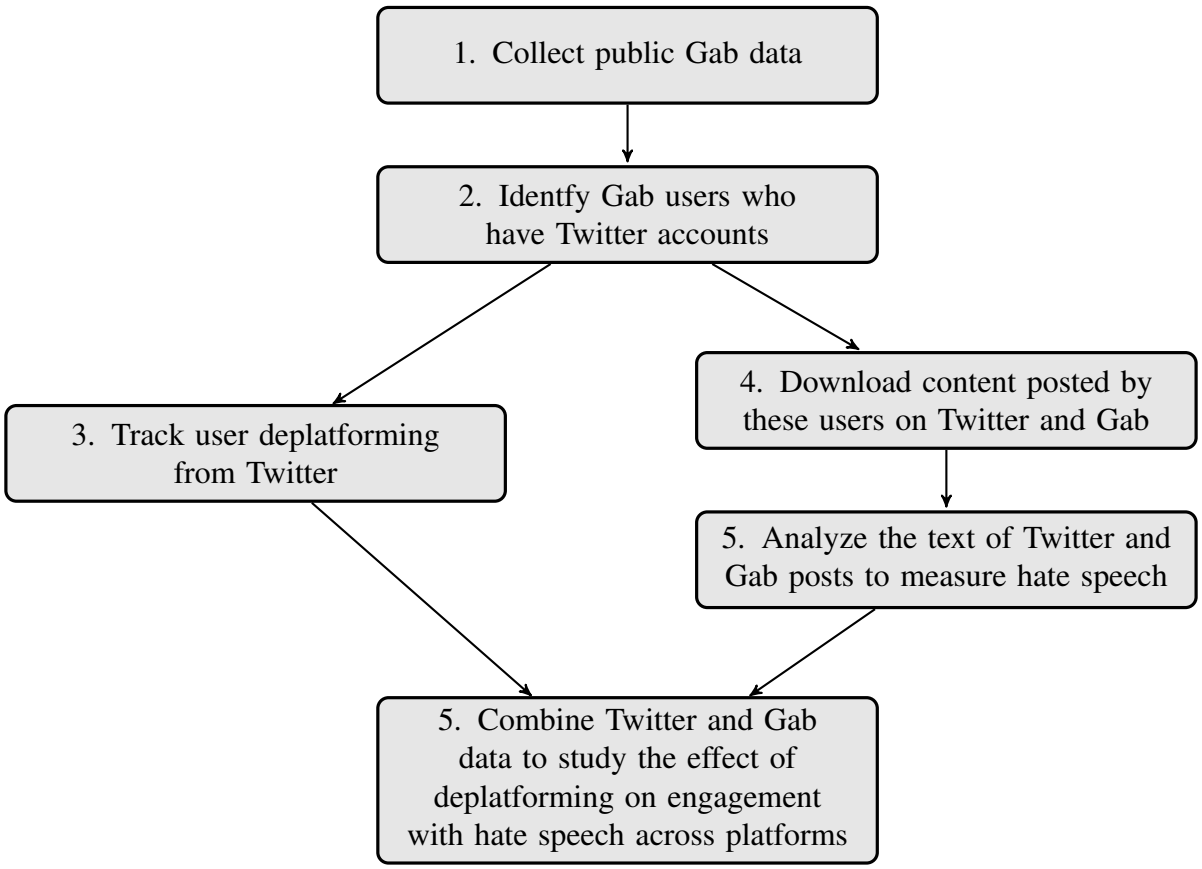
4.1 Collecting Gab data

Gab is a social media platform popular among far-right communities in the United States. In recent years, with the rise in content moderation by mainstream platforms and the suspension of accounts promoting hate speech, Gab has seen a dramatic growth in its user base, especially among those on the fringes. Today, the platform is widely used by various extremist communities, including those affiliated with Neo-Nazi, white supremacy, and other far-right groups, as well as communities advocating conspiracy theories like QAnon. Gab promotes itself as "the home of free speech online,"[4] primarily because it engages in almost no content moderation.

The spread of hate speech on Gab prompted widespread criticism of the platform. Critics accuse Gab of providing a safe haven for hate groups, which can facilitate radicalization and offline

[4] `https://gab.com/about`

Figure 2: **Data Collection Overview**



harm. In October 2018, after the Pittsburgh synagogue shooting, Gab was banned from the servers of its hosting provider, after it was found that the perpetrator was an active Gab user who publicized his attack on the platform.[5] Gab was also among the platforms that were actively used to promote the storming of the U.S. Capitol on January 6, 2021, which resulted in the death of five people and the injury of many.[6]

To collect Gab data, my first step was to download all posts that were publicly available on the platform, without any filtering. Using the API of the Mastodon social network, on which Gab has

[5] https://www.reuters.com/article/us-pennsylvania-shooting-gab/gab-com-goes-offline-after-pittsburgh-synagogue-shooting-idUSKCN1N20Q5

[6] https://www.npr.org/sections/insurrection-at-the-capitol/2021/01/07/954671745/on-far-right-websites-plans-to-storm-capitol-were-made-in-plain-sight

been operating since July 2019,[7] I collected information on all posts that were viewable on Gab's public timeline. This initial collection allowed me to obtain data on Gab users who were active on the platform since October 2019. In June 2020, Gab disabled the public timeline. To continue collecting data, I moved to a retrospective collection in which I obtained all historical posts by Gab users who were in my dataset by that point, i.e., who had made a public post prior to June 2020. After January 2021, I was again able to access Gab's public timeline, which allowed me to resume both prospective and retrospective data collection. The total number of unique Gab users in my sample is 93,004.

4.2 Identifying Gab users who have Twitter accounts

My second step was to identify Gab users who have Twitter accounts. My approach was simple: I queried the Twitter API for screen names used by Gab users. Matching screen names is a popular approach to find user linkages across platforms. There are other methods that one can use, some of which draw on the similarity in the content that users post across platforms, as well as similarity in user metadata (profile pictures, profile descriptions, etc.) (Shu et al., 2017; Hadgu and Gundam, 2019). While matching screen names likely undercounts users who do not use the same name across platforms, this approach is less likely to include accounts of unrelated users. Using the matched screen name approach, I identified 30,444 Gab users who had active Twitter accounts (33% of the Gab users in my sample). Manual validation of the matched accounts confirmed that these are owned by the same individuals. Figure 3 provides some examples.

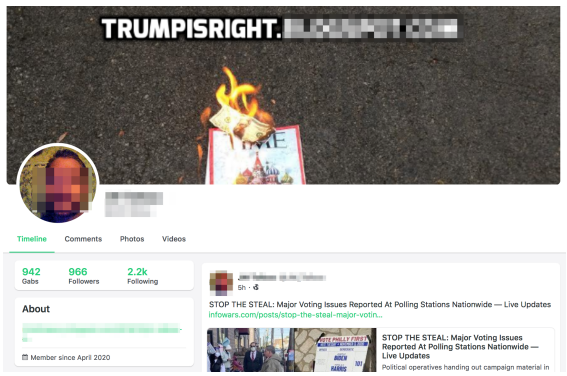
4.3 Quantifying hate speech

The next step was to measure the extent to which users on both platforms engage with hate speech. Since the volume of Gab and Twitter posts is high, I draw on automated text analysis methods to

[7] 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`.

Figure 3: Twitter and Gab Accounts with Identical Screen Names

(A) Gab account



(B) Twitter account

(C) Gab account

(D) Twitter account

Note: The Figure presents Gab and Twitter profiles of the same individuals. The profiles were matched via screen names.

measure hate speech. While definitions of hate speech vary, many capture a similar concept. In this study, I use the description of Encyclopedia of Political Communication to define online hate speech. According to this definition, hate speech consists of:

> *"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." (Kaid and Holtz-Bacha, 2007)

Using several machine learning models, I 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, I measured two related topics that are popular among far-right communities: endorsement of white supremacy and general discourse on gun policy in the United States.

Figure 4 illustrates the text analysis process. First, I randomly sampled 124K Twitter and Gab posts from the full post-level data that I downloaded from both platforms. Second, I had a large team of research assistants label this sample along the categories listed above. The Appendix lists the categories and their definitions. Third, I used the labeled dataset to train Naive Bayes classifiers – one for each category – to detect hate speech in the unlabeled posts. Since class imbalance was high in the training set, I rebalaned the training data by over-sampling posts from the minority category. I used 80% of the labeled data for training and evaluated the models' performance on the remaining 20%. As Appendix Table A1 shows, out-of-sample performance was high, where the accuracy, prediction, and recall were above 0.9.

Figure 4: Detecting Hate Speech on Gab and Twitter

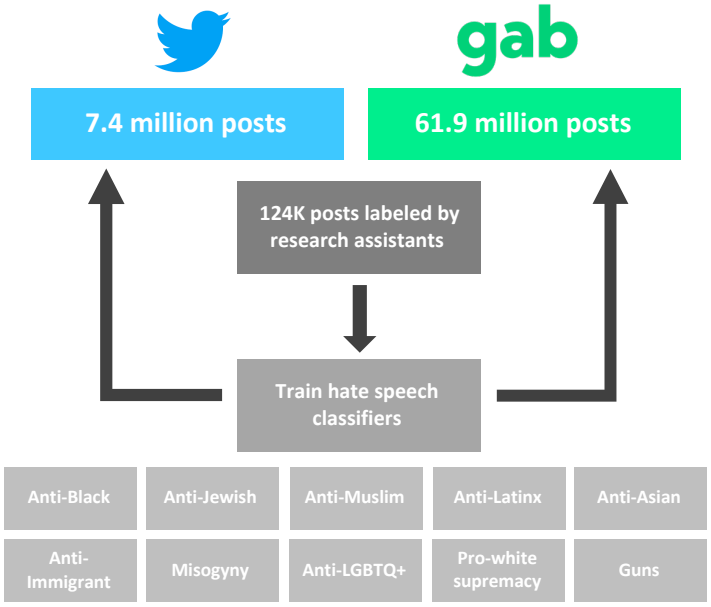


Table 1 shows the words that were identified by the models to be predictive of each hate speech category. Examining these words is useful for understanding the themes that were captured in each topic. As expected, much of the predictive content includes words that refer to the targeted groups. But a close look shows particular themes that characterize hate speech along these categories. 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.

Table 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
Guns	gun, people, die, police, government, time, control, shot, state, man, black, amendment, cops, video, firearms, weapons

Table 2: Summary Statistics

A. Twitter

	N	Mean	St. Dev.
Hate speech index*	7,440,991	1.062	1.405
Hateful content	7,440,991	0.231	0.272
Anti-Black	7,440,991	0.151	0.255
Anti-Jewish	7,440,991	0.095	0.210
Anti-Asian	7,440,991	0.084	0.191
Anti-Muslim	7,440,991	0.078	0.189
Anti-Latinx	7,440,991	0.035	0.142
Anti-Immigrant	7,440,991	0.086	0.197
Misogyny	7,440,991	0.092	0.216
Anti-LGBTQ+	7,440,991	0.097	0.216
Pro-white supremacy	7,440,991	0.113	0.212
Guns	7,440,991	0.100	0.199

B. Gab (matched sample)

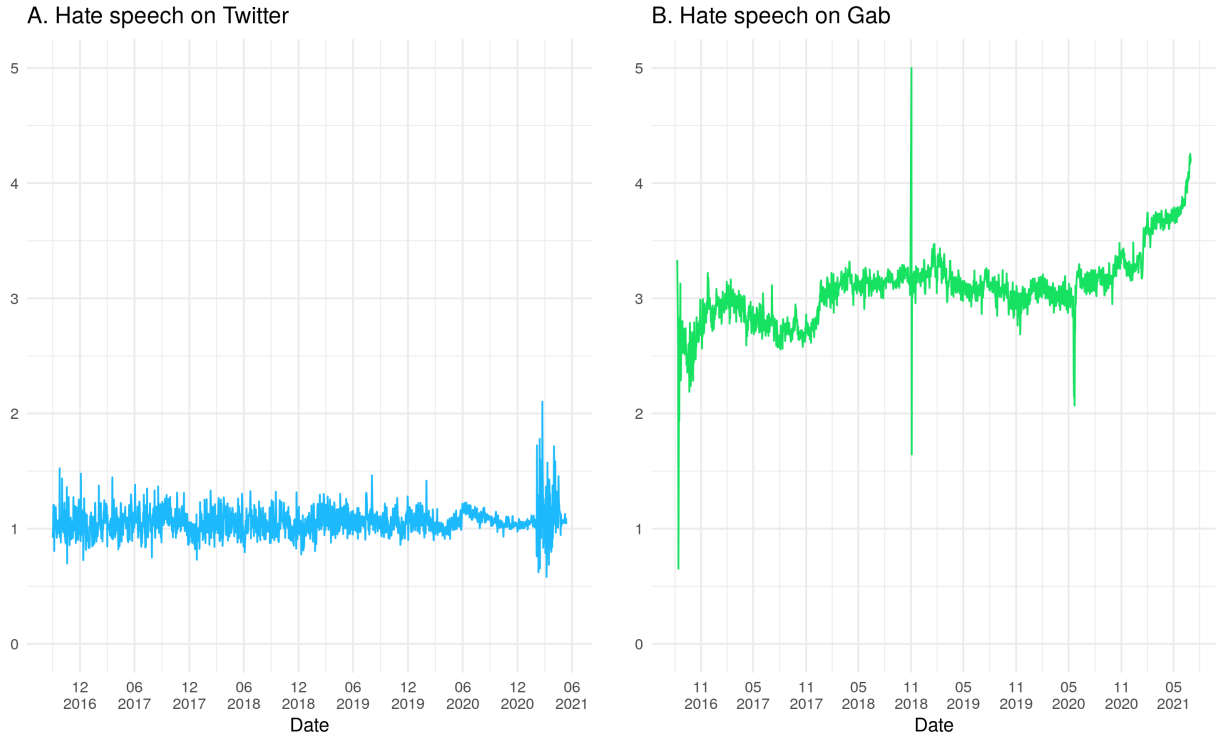
	N	Mean	St. Dev.
Hate speech index*	18,935,330	3.131	2.154
Hateful language	18,935,330	0.412	0.263
Anti-Black	18,935,330	0.330	0.264
Anti-Jewish	18,935,330	0.319	0.265
Anti-Asian	18,935,330	0.287	0.253
Anti-Muslim	18,935,330	0.299	0.259
Anti-Latinx	18,935,330	0.260	0.256
Anti-Immigrant	18,935,330	0.299	0.259
Misogyny	18,935,330	0.298	0.263
Anti-LGBTQ+	18,935,330	0.305	0.263
Pro-white supremacy	18,935,330	0.323	0.259
Guns	18,935,330	0.305	0.252

Note: Hate speech index captures content from all of the categories, except guns. It range between 0 and 10.

Table 2 shows summary statistics of the hate speech variables for both datasets. Panel A shows summary statistics for the Twitter data, and Panel B presents this information for the sample of the Gab users who also have Twitter accounts. Appendix Table A2 shows that the distributions of hate speech are similar in the full Gab data, which also includes users without Twitter accounts. Overall, tables show that users posted hateful content on both platforms, but the level of hate speech in Gab is higher. In much of the analysis presented below, I'll be using a hate speech index that ranges between 0 and 10 and captures content from all categories except guns.

Figure 5 shows the average level of hate speech on both platforms over time. While the level of hate speech on Twitter is fairly constant, on Gab there is more temporal variation. In particular, the Gab data shows a rise and then drop in hate speech at the end of 2018, which is driven by the banning of the platform from its hosting provider after the Pittsburgh synagogue shooting. After a week of limited service, Gab came back online – this time, running on its own independent servers (Robertson, 2018). Since then, hate speech on Gab has been on the rise, and significantly increased after summer 2020, when the Black Lives Matter protests took place across the United States, as well as after the attack on the U.S. Capitol in January 2021.

Figure 5: Hate speech on Twitter and Gab Over Time



Note: The Figure presents the daily average of hate speech posts on Twitter Gab between 2016 and 2021.

4.4 Measuring deplatforming

The main independent variable in this study is user suspension from Twitter. My goal is to examine how deplatforming from mainstream platforms affects user behavior on alternative platforms. As mentioned above, current literature on deplatforming focuses almost exclusively on within-platform effects, with little attention given to the consequences of deplatforming across platforms.

To track user suspension from Twitter, I queried the Twitter API for the profile metadata of the 30,444 Twitter users who also have Gab accounts. When a user is suspended from Twitter, the query returns no data for their account. I used this information to prospectively track deplatforming between September 2020 and February 2021. To ensure that the accounts that I identified as suspended were indeed banned from the platform, I had research assistants manually validate the list of suspended accounts by checking their user status on Twitter on July 2021. Overall, 2,200 of the users in my sample were suspended during this period. Figure 6 shows the daily number

of suspensions. While deplatforming events took place every day, there were two large peaks in the number of suspensions in late September 2020 and early January 2021, which relate to broader campaigns by Twitter that targeted account promoting the QAnon conspiracy theory and and the January 6 storming of the U.S. Capitol.

To further validate my suspension measure, I examined whether suspended users talked about their suspension on Gab. The idea is that if suspension drives users to alternative platforms, we should see discourse on the deplatforming among Gab users who were banned from Twitter. In addition, if mentions of Twitter are driven by the deplatforming, we should see an increase in discourse on Twitter especially among deplatformed users – non-deplatformed users should not talk more about Twitter on these exact dates.

Figure 7 plots mentions of Twitter in Gab posts by suspended (blue) and non-suspended users (orange). The figure normalizes the difference in days between the suspension date and the timing of Gab posts. Consistent with the deplatforming interpretation, I find that mentions of Twitter on Gab significantly increase among suspended users on the day of suspension, but do not show any time trend for non-suspnded users. Table 3 shows some examples of deplatformed users' posts on Gab. These posts clearly relate to their suspension from Twitter. To further ensure that the increase in Twitter mentions is not spurious, I also examined mentioning of other mainstream platforms like Facebook and YouTube on the dates of Twitter's deplatforming. Appendix Figure A1 shows a different pattern in the mentions of Facebook or YouTube around these dates.

4.5 Can we treat the timing of deplatforming as a plausibly exogenous shock?

In order to interpret the link between deplatforming and subsequent changes in online behavior as causal, the timing of suspension from Twitter would need to be exogenous to users' radicalization process. This, however, might not be the case if the particular instance when a user is deplatformed is endogenous to the content that they post on Twitter. For example, it might be that Twitter suspends accounts after they cross a pre-defined threshold determined by the platform's content moderation policies. If we observe users expressing more hate on Gab after being deplatfomed

Figure 6: Twitter Suspensions by Date

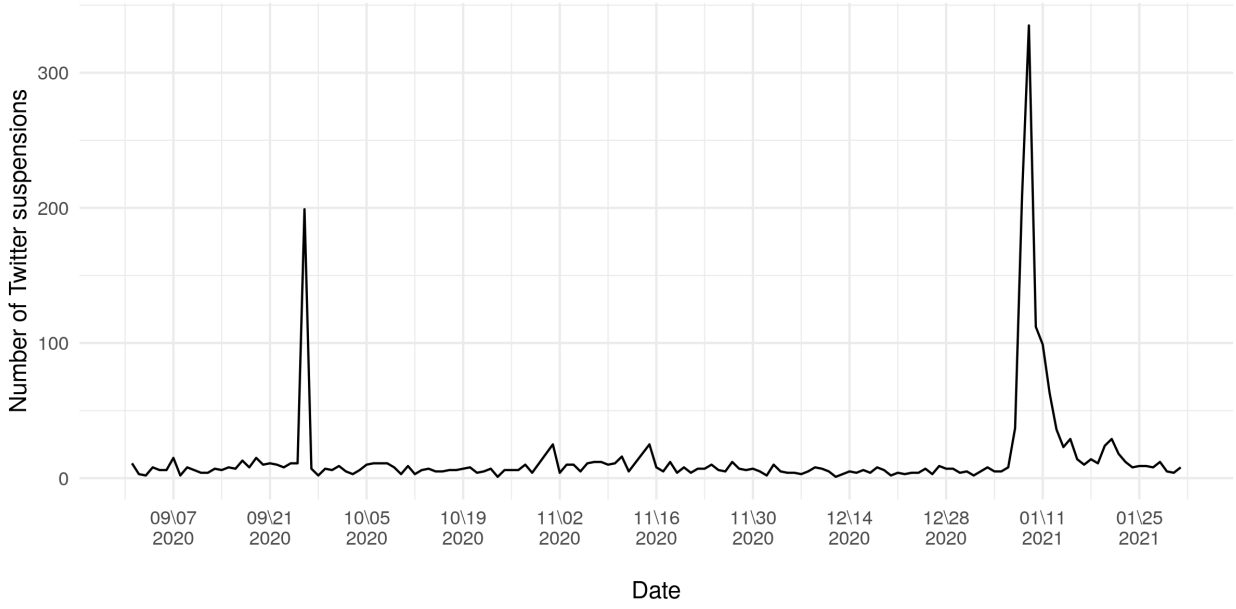
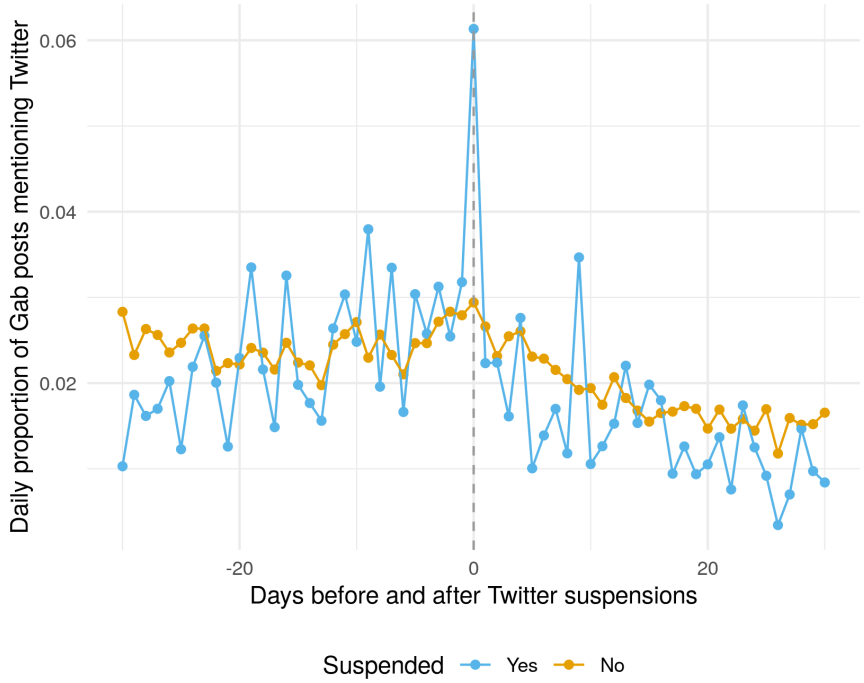


Figure 7: Posts Mentioning Twitter on Gab



from Twitter, this may have nothing to do with suspension, but simply a reflection of a time trend in users' engagement with hate speech that would have taken place regardless of suspension. This

Table 3: Posts Mentioning Twitter by Deplatformed Users on Gab

1	*I really need to use gab more often I'm fed up with a constant bans on Twitter just for telling the truth...*
2	*F** Twitter and their censorship of harmful words*
3	*Twittersux i said Trump is the right man for America. They suspended me*
4	*So my Twitter account has been blocked for (apparently) disseminating the NY Post story on Hunter Biden. L'il ole me... So I'm joining our Press Secretary and many others blocked by Jack Dorsey for fighting for the right to free speech. Hello Gab!*
5	*the twitter jews got me*
6	*I hate Twitter.*
7	*My twitter is suspended for talking too much about #Plandemic*
8	*I've about had it with @Jack and Twitter censoring free speech. I joined Gab a while back but haven't used it much. I think that's about to change!*

possibility is reflected in the conceptual graph in Figure 8A.

However, the timing of suspension may not be so deterministic. There is growing evidence suggesting that the particular moment in which deplatforming happens is arbitrary, at least within a particular window. This is partly driven by the fact that the rules that determine what is allowed on the platform are not fixed, but change over time. For example, content promoting the QAnon conspiracy theory was allowed to widely circulate on mainstream social media until summer 2020, after which companies began taking enforcement actions against such content. Twitter did not suspend QAnon-promoting accounts until July 2020,[8] and Facebook took similar actions against groups, pages, and posts linked to QAnon only after August 2020.[9] Because of this 'moving target' of content moderation, a lot of content ends up in a gray zone: even though it could be violating the company's rules, it is not actively taken down. In fact, there is some evidence that extremist activists began exploiting content moderation ambiguities to promote their agendas on mainstream social media by disseminating content that is close to, but doesn't officially break the rules.[10]

As a result, the particular moment in which an account is deplatformed might be plausibly exogenous to a user's radicalization path, at least around the time of the suspension. This is reflected

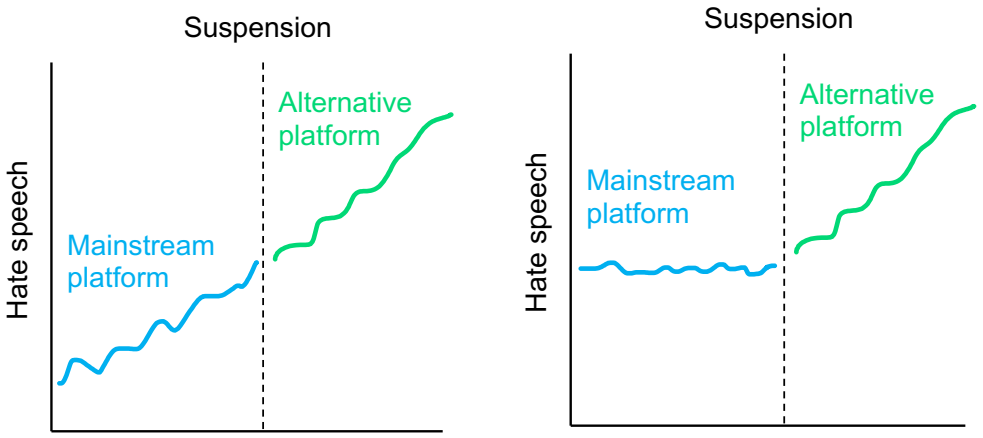
[8] `https://twitter.com/TwitterSafety/status/1285726277719199746?s=20`

[9] `https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/`

[10] See: `https://www.washingtonpost.com/technology/2021/03/14/facebook-vaccine-hesistancy-qanon/`

in the conceptual graph in Figure 8B, which shows no clear pattern in hate speech before suspension. In this case, post-deplatforming changes in hate speech on alternative social media might be driven by the act of suspension itself.

Figure 8: Conceptual Graphs of Suspension and Hate Speech



(A) Endogenous suspension timing

(B) Exogenous suspension timing

To empirically examine this possibility, in Figure 9 I plot hate speech posted on Twitter in the two months before suspension. Since Twitter suspended users in different dates, I center the data around a normalized suspension day. The figure presents the daily number of hate speech tweets, measured with the hate speech index, for suspended and non-suspended users.[11] Since I have many more non-suspended users than suspended users in my dataset, the overall number of hate speech tweets is higher for the non-suspended group. However, as the figure shows, there does not seem to be a clear time trend that drives the suspension. I find similar patterns in Figure 10, which shows the data for each category separately. In all cases, the time trends in hate speech on Twitter do not seem to be correlated with suspension.

[11] To count hate speech tweets by day and suspension status, I created a binary version of the index that codes tweets as including hate speech if their index value is above the mean and zero otherwise.

Figure 9: Pre-suspension Trends in Hate Speech on Twitter

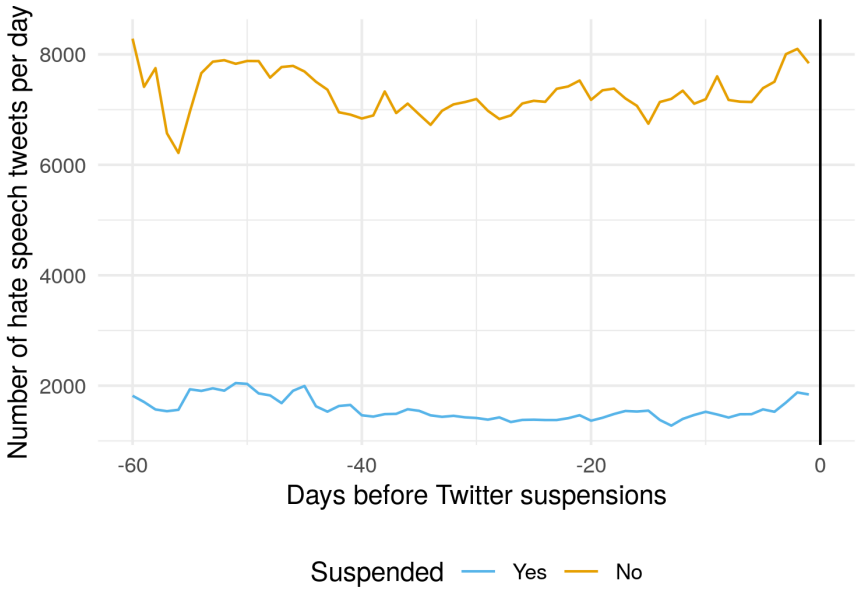
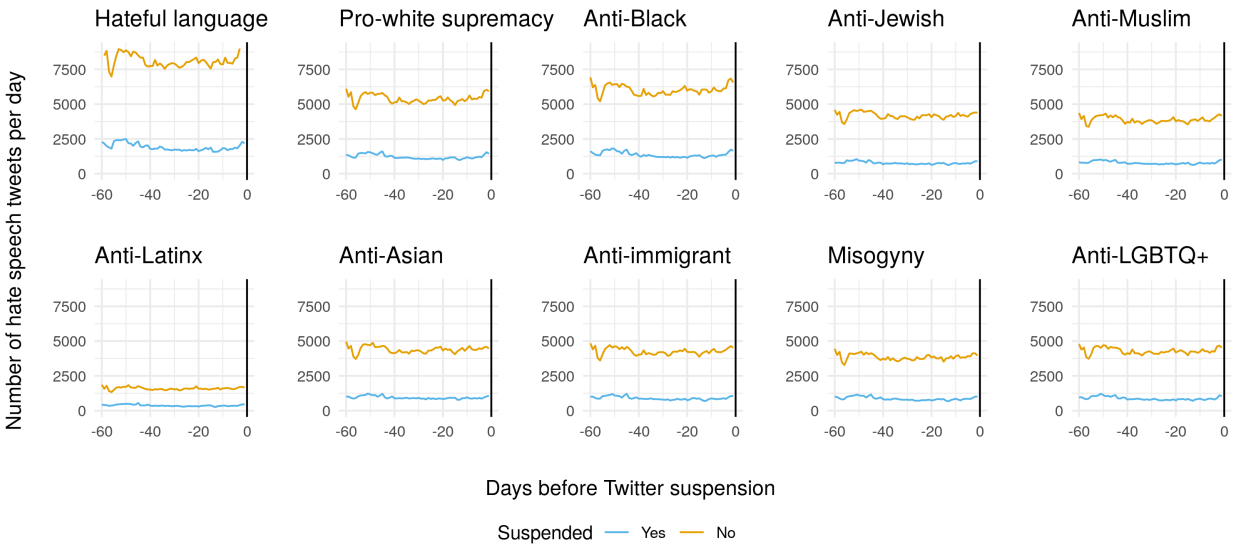


Figure 10: Pre-suspension Trends in Hate Speech on Twitter (By Category)



5 Research Design

Drawing on the findings from the previous section, I designed a study that examines the relationship between Twitter deplatforming and engagement with hate speech on Gab. I begin with a simple descriptive analysis that compares the level of hate speech by banned users before and after suspension. Out of the 2,200 users in my sample that were suspended from Twitter, 745 actively posted on Gab during the months of the study. I focus on these users here. If deplatforming mobilizes hate on alternative platforms, we should see an increase in hate speech expressed by banned users on Gab after their suspension from Twitter. I generated an aggregated variable that measures the average level of hate speech in each user's Gab posts before and after deplatforming events. I then estimated the following model:

$$\text{Hate speech}_{it} = \beta(\text{Post-deplatforming}_{it}) + \varepsilon_i$$

Hate speech is the value of the hate speech index for user i in time period t, and *Post-deplatforming* indicates whether the content was posted before or after the user's suspension from Twitter. Standard errors are clustered at the user level. Since I do not have a strong prior on the time window around which we should see changes in hate speech (is it immediate? does it take place over time?), I examine the results for different windows, ranging from one to ninety days after deplatforming.

However, a simple over-time comparison might lead to inaccurate inferences, especially if other factors drive overtime changes in hate speech. Since the data collection took place during a very heated time period, especially around the U.S. elections and the January 6 storming of the Capitol, it is very likely that hateful content has been increasing on Gab regardless of deplatforming. To control for these trends, I estimate additional difference-in-differences models that take into account temporal changes in hate speech, by comparing users who are deplatformed to those that are not. The β_3 coefficient in the model below captures post-deplatforming changes in hate speech among deplatformed users, after differencing out temporal changes in hate speech in the

rest of the sample:

$$\text{Hate speech}_{it} = \beta_1(\text{Post deplatforming}_{it}) + \beta_2(\text{Deplatformed user}_i) +$$

$$\beta_3(\text{Post deplatforming}_{it} \times \text{Deplatformed user}_i) + \varepsilon_i$$

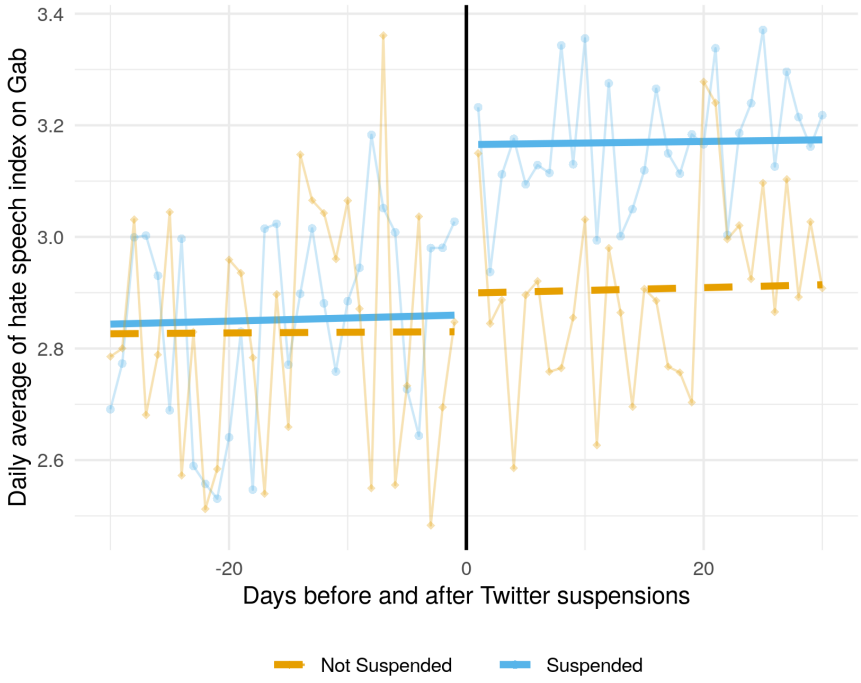
I estimate this model on two different samples. First, I compare users who are banned from Twitter to those that are not banned but have a Twitter account. In this analysis, all Gab users who have a Twitter account but are not deplatformed serve as a counterfactual to Gab users who are suspended from Twitter. I randomly assign placebo suspension dates to the non-suspended sample, and examine changes in hate speech around deplatforming events.

Since there could be important differences between these groups—for example, suspended and non-suspended users express different levels of hate speech on Twitter (see Figures 9 and 10)—I also examine a second sample that only compares users that are matched on the level of hate speech before suspension. To match users, I created an aggregate measure of hate speech that each user posted on Gab and identified a comparison case for each deplatformed user with nearest-neighbor matching.[12]

Figure 11 plots the overtime trends in hate speech by deplatformed users (blue) and their matched sample (orange). The x-axis is the number of days between a deplatforming event and the date in which these users posted on Gab. In order to simultaneously observe trends for all deplatforming events, I normalized the difference in days between the events and the timing of Gab posts. The figure shows that the over-time trends in hate speech are parallel in the pre-treatment period. Only after deplatforming we observe a shift in those trends, where hate speech by deplatformed users increases, but the rhetoric of non-deplatformed users remains the same.

[12]Daniel E. Ho, Kosuke Imai, Gary King, Elizabeth A. Stuart (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Vol. 42, No. 8, pp. 1-28. URL: `http://www.jstatsoft.org/v42/i08/`

Figure 11: Twitter Suspensions and Hate Speech on Gab



6 Results

Table 4 presents findings from a simple comparison of hate speech by Gab users who were banned from Twitter before and after their suspension. The table shows the relationship between deplatforming and hate speech across various post-deplatforming windows, where the hate speech index is the dependent variable. Gab users who were suspended from Twitter significantly increased their hate speech in the weeks and months after their suspension. Within 90 days of their Twitter ban, the level of hate expressed by deplatformed users on Gab was about 17% higher than the pre-suspension period.

Next, I examine the results of the difference-in-differences analysis. Since the level of hate speech on Gab was on the rise during the period of the study (see Figure 5), the patterns shown in Table 4 could be driven factors other than deplatforming. To account for these time trends, I examine whether the change in hate speech by suspended users after deplatforming is different from the change in hate speech by non-suspended users during the same time period. The DiD coeffi-

Table 4: Hate speech on Gab after Twitter Deplatforming

Post window	1 day	2 days	7 days	14 days	30 days	60 days	90 days
Post	0.608***	0.449***	0.491***	0.513***	0.543***	0.495***	0.452***
	(0.168)	(0.146)	(0.116)	(0.108)	(0.105)	(0.106)	(0.110)
Constant	2.624***	2.624***	2.624***	2.624***	2.624***	2.624***	2.624***
	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)	(0.114)
Observations	49,411	49,532	50,101	50,953	52,780	55,298	56,890
Clusters	733	733	734	734	738	745	745
R^2	0.0002	0.0003	0.001	0.002	0.005	0.007	0.007
Note:							*p<0.1; **p<0.05; ***p<0.01

cients presented in Table 5 show a similar pattern: hate speech among suspended users increases in the post-deplatforming period, even after accounting for overtime changes in hate speech by non-deplatformed users. The magnitude of the DiD coefficient diminishes after 60 days, partly because there is an overall increase in hate-speech among non-deplatformed users during these windows.

Table 5: Hate speech on Gab after Twitter Deplatforming (DiD, Full Sample)

Post window	1 day	2 days	7 days	14 days	30 days	60 days	90 days
Deplatformed	−0.015	−0.015	−0.015	−0.015	−0.015	−0.015	−0.015
	(0.116)	(0.116)	(0.116)	(0.116)	(0.116)	(0.116)	(0.116)
Post	0.255***	0.212***	0.232***	0.276***	0.289***	0.297***	0.301***
	(0.037)	(0.030)	(0.023)	(0.021)	(0.020)	(0.020)	(0.020)
Deplatformed × Post	0.354**	0.237	0.259**	0.237**	0.255**	0.198*	0.151
	(0.172)	(0.148)	(0.119)	(0.110)	(0.107)	(0.108)	(0.111)
Constant	2.639***	2.639***	2.639***	2.639***	2.639***	2.639***	2.639***
	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)	(0.019)
Observations	1,728,930	1,731,263	1,742,695	1,759,545	1,796,540	1,853,523	1,896,415
Clusters	30,234	30,234	30,253	30,271	30,357	30,434	30,440
R^2	0.00002	0.00003	0.0002	0.0004	0.001	0.002	0.002
Note:							*p<0.1; **p<0.05; ***p<0.01

In Table 6 I replicate the analysis for a sample of Gab users who are matched on the basis of hate speech posted on Gab before suspension. Here, the counterfactual group consists of Gab users who posted similar content to suspended users before the latter were deplatformed. I find a similar pattern, where engagement with hate speech increases after suspension. In this sample, however, the magnitude of the DiD coefficients is smaller in the 1-day and 2-day windows, and is not statistically significant. As in the previous analyses, I find that hate speech significantly rises

among deplatformed users several weeks after suspension. This might suggest that radicalization on fringe platforms in the post-suspension period is not immediate, but takes place after individuals spend time in social networks active in promoting hateful content and extremist views.

To get a better sense of the types of hate speech that increase after deplatforming, I break the hate speech index into its constituent categories in Table 7. Each column in the table reports the DiD results, using the matched sample, for a different hate speech category.[13] The table presents findings from the 30-day window, but the results remain similar when using other windows as well. I find that Twitter deplatforming induced more hate speech on Gab targeting many different identity groups — a change that is not present for non-deplatformed Gab users. When comparing the post-suspension level of hate speech by banned users to their pre-suspension levels, the change reflects an increase of about 20% in most categories. Beyond hateful language, I find that deplatforming also increased banned users' engagement with content endorsing white supremacy. This category consists of a lot of content promoting the QAnon conspiracy theory and white nationalism more generally (see Table 1).

Table 6: Hate speech on Gab after Twitter Deplatforming (DiD, Matched Sample)

Post window	1 day	2 days	7 days	14 days	30 days	60 days	90 days
Deplatformed	−0.193	−0.193	−0.193	−0.193	−0.193	−0.193	−0.193
	(0.156)	(0.156)	(0.156)	(0.156)	(0.156)	(0.156)	(0.156)
Post	0.332	0.176	0.049	0.033	0.090	0.126	0.153
	(0.242)	(0.186)	(0.149)	(0.137)	(0.133)	(0.122)	(0.123)
Deplatformed × Post	0.276	0.274	0.437**	0.480***	0.454***	0.369**	0.298*
	(0.295)	(0.236)	(0.189)	(0.175)	(0.170)	(0.162)	(0.165)
Constant	2.817***	2.817***	2.817***	2.817***	2.817***	2.817***	2.817***
	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)	(0.107)
Observations	92,699	92,878	93,719	94,997	97,695	101,572	104,181
Clusters	1,467	1,467	1,467	1,468	1,469	1,470	1,470
R^2	0.003	0.003	0.003	0.004	0.005	0.006	0.006
Note:							*p<0.1; **p<0.05; ***p<0.01

[13]The outcome measures are binary variables that are coded 1 when the posterior probability estimated by the Naive Bayes models was greater than the average for each category, and 0 otherwise. The results hold when using the predicted probability as the outcome.

Table 7: DiD Estimates by Hate Category (Matched Sample)

	Hateful language	**Pro-white supremacy**	**Anti-Black**	**Anti-Jewish**	**Anti-Muslim**
Deplatformed	−0.036	−0.043	−0.035	−0.037	−0.031
	(0.029)	(0.031)	(0.033)	(0.033)	(0.034)
Post	0.017	0.022	0.026	0.019	0.023
	(0.026)	(0.027)	(0.028)	(0.029)	(0.029)
Deplatformed × Post	0.076**	0.094***	0.088**	0.099***	0.095**
	(0.032)	(0.035)	(0.036)	(0.036)	(0.037)
Constant	0.607***	0.543***	0.537***	0.523***	0.496***
	(0.020)	(0.020)	(0.022)	(0.023)	(0.024)
R^2	0.003	0.005	0.004	0.004	0.004
% Change	*13.4*	*18.7*	*17.6*	*20.4*	*20.5*
	Anti-Latinx	**Anti-Asian**	**Anti-immigrant**	**Misogyny**	**Anti-LGBTQ+**
Deplatformed	−0.021	−0.025	−0.030	−0.030	−0.032
	(0.035)	(0.034)	(0.034)	(0.035)	(0.035)
Post	0.035	0.031	0.021	0.038	0.026
	(0.031)	(0.029)	(0.029)	(0.029)	(0.029)
Delatformed × Post	0.088**	0.087**	0.090**	0.090**	0.092**
	(0.039)	(0.037)	(0.037)	(0.037)	(0.037)
Constant	0.445***	0.486***	0.500***	0.496***	0.505***
	(0.025)	(0.024)	(0.023)	(0.024)	(0.024)
R^2	0.003	0.003	0.003	0.004	0.004
% Change	*20.7*	*19.0*	*19.3*	*19.3*	*19.5*
Observations	97,539	97,539	97,539	97,539	97,539
Clusters	1,469	1,469	1,469	1,469	1,469

Note: *p<0.1; **p<0.05; ***p<0.01

7 Discussion and Conclusion

What do these results mean for deplatforming more generally? This study shows that it's important to take into account cross-platform dynamics when evaluating content moderation policies. As mainstream platforms expand their enforcement actions against hate groups operating on their sites, it is important to keep an eye on alternative social media platforms that are becoming viable alternatives. Gab is just one of many new sites that have popped up in recent years. Future research on online extremism could examine whether similar dynamics are happening on other social media platforms that welcome banned users.

My main finding is that deplatformed users express more hate speech on alternative platforms in the months following their ban from mainstream platforms. However, I did not explore the mechanisms that might drive this change. Do banned users deepen their engagement with hate speech because of the content to which they are exposed on fringe platforms? If so, what content is most likely to radicalize? Prior research shows that extremists' online campaigns are not always effective, and that certain types of content are more likely to increase engagement than others (Mitts, Phillips and Walter, 2021). A fruitful direction for future work is to study online campaigns targeting deplatformed communities on fringe platforms, in order to examine their effectiveness in radicalizing potential supporters.

More broadly, this study points to the potential benefits of cross-platform collaboration. If content moderation encourages migration to other platforms, then coordination between platforms—both in terms of moderation policies as well as in terms of enforcement—could help mitigate the problem of radicalization on the fringes. Recent efforts by the Global Internet Forum to Counter Terrorism, which has been leading cross-platform initiatives in the fight against extremism, provide an example of how this could be done.[14] By facilitating data-sharing across platforms, it enables early detection and takedown of violating content by multiple social media platforms simultaneously.

[14] `https://gifct.org`

However, even if such collaboration continues, technology will always provide new places for extremists to gather. While this study focused on content moderation and deplatforming, a new wave of research suggests that taking down content and banning accounts might not be the only way to combat hate and extremism. Counter-speech interventions, which directly engage those who express hate online, might be an effective tool to reduce hate speech on mainstream platforms as well (Munger, 2017; Davey, Birdwell and Skellett, 2018; Siegel and Badaan, 2020). Either way, better understanding the ways in which online interventions shape extremists' activity on social media is likely to remain an important research question in the years to come.

References

Beach, Stephanie J. 2019. "Hashtag Hate: The Need for Regulating Malignant Rhetoric Online." *Vt. L. Rev.* 44:129.

Bennett, Tom. 2018. "Gab Is the Alt-Right Social Network Racists Are Moving to." *Vice* .

Chandrasekharan, Eshwar, 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." *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW):1–22.

Cofnas, Nathan. 2019. "Deplatforming Won't Work." *Quillette* .

Conger, Kate. 2021. "Twitter, in Widening Crackdown, Removes Over 70,000 QAnon Accounts." *The New York Times* .

Davey, Jacob, Jonathan Birdwell and Rebecca Skellett. 2018. "Counter Conversations: A model for direct engagement with individuals showing signs of radicalisation online." *Institute for Strategic Dialogue* .

Enders, Walter and Todd Sandler. 2011. *The political economy of terrorism*. Cambridge University Press.

Frenkel, Sheera. 2021. "The storming of Capitol Hill was organized on social media." *The New York Times* .

Greer, Ryan. 2020. "Weighing the Value and Risks of Deplatforming." *The New York Times* .

Hadgu, Asmelash Teka and Jayanth Kumar Reddy Gundam. 2019. User Identity Linking Across Social Networks by Jointly Modeling Heterogeneous Data with Deep Learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. pp. 293–294.

Heilweil, Rebecca and Shirin Ghaffary. 2021. "How Trump's internet built and broadcast the Capitol insurrection: Online extremists started planning the chaos of January 6 months ago." *Global Network on Extremism and Technology* .

Kaid, Lynda Lee and Christina Holtz-Bacha. 2007. *Encyclopedia of political communication*. SAGE publications.

Kydd, Andrew H and Barbara F Walter. 2006. "The strategies of terrorism." *International security* 31(1):49–80.

Mitts, Tamar, Gregoire Phillips and Barbara Walter. 2021. "Studying the Impact of ISIS Propaganda Campaigns." Forthcoming in the *Journal of Politics*.

Munger, Kevin. 2017. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior* 39(3):629–649.

Nouri, Lella, Nuria Lorenzo-Dus and Amy-Louise Watkin. 2019. "Following the Whack-a-Mole: Britain First's Visual Strategy from Facebook to Gab." *Global Research Network on Terrorism and Technology Paper* (4).

Persily, Nathaniel and Joshua A Tucker. 2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*. Cambridge University Press.

Roberts, Margaret E. 2018. *Censored*. Princeton University Press.

Robertson, Adi. 2018. "Gab is back online after being banned by GoDaddy, PayPal, and more." *The Verge* .

URL: *https://www.theverge.com/2018/11/5/18049132/gab-social-network-online-synagogue-shooting-deplatforming-return-godaddy-paypal-stripe-ban*

Romm, Tony. 2021. "Facebook, Twitter could face punishing regulation for their role in U.S. Capitol riot, Democrats say." *The Washington Post* .

Schroeder, Jared. 2018. "Toward a discursive marketplace of ideas: Reimaging the marketplace metaphor in the era of social media, fake news, and artificial intelligence." *First Amendment Studies* 52(1-2):38–60.

Shu, Kai, Suhang Wang, Jiliang Tang, Reza Zafarani and Huan Liu. 2017. "User identity linkage across online social networks: A review." *Acm Sigkdd Explorations Newsletter* 18(2):5–17.

Siegel, Alexandra A and Vivienne Badaan. 2020. "# No2Sectarianism: Experimental approaches to reducing sectarian hate speech online." *American Political Science Review* 114(3):837–855.

Smith, Aaron. 2018. "Public Attitudes Toward Technology Companies." *Pew Research Center* . **URL:** *https://www.pewresearch.org/internet/2018/06/28/public-attitudes-toward-technology-companies/*

Sullivan, mark. 2021. "Facebook has deleted 19,500 groups tied to 'militarized social movements'." *Fast Company* .

Thomas, Daniel and Laila Wahedi. 2021. "Disrupting Hate: The Effect of Deplatforming Hate Organizations on their Online Audience." Unpublished working paper.

Thompson, Peter A. 2019. "Beware of geeks bearing gifts: Assessing the regulatory response to the Christchurch Call." *The Political Economy of Communication* 7(1).

8 Appendix

8.1 Label definitions for hate speech training set

1. **Hateful language**: posts expressing prejudice and contempt toward individuals from a non-white social groups, including pejoratives and group-based insults that can appear as negative labels or narratives about the group's alleged negative behavior (source: Encyclopedia of Political Communication, Sage, 2007)
2. **Anti-Black**: posts expressing hate towards individuals from black and/or African American background. Includes comments criticizing inter-racial marriage.
3. **Anti-Jewish**: posts expressing hate towards individuals from a Jewish background. Includes comments criticizing marriage between whites and Jews.
4. **Anti-Muslim**: posts expressing hate towards individuals from a Muslim background. Includes comments criticizing marriage between whites and Muslims.
5. **Anti-Asian**: posts expressing hate towards individuals from an Asian, East Asian, or South-Asian background. Includes comments criticizing white and Asian marriage.
6. **Anti-Latinx**: posts expressing hate towards individuals from a Latin American background. Includes comments criticizing marriage between whites and Latinos/Latinas.
7. **Anti-immigrant**: posts expressing hate towards immigrants more generally, including refugees, asylum seekers, and other types of migrants. Includes posts criticizing open borders and liberal immigration policies.
8. **Misogyny**: posts conveying misogyny and hate speech against women.
9. **Anti-LGBTQ+**: posts expressing hate towards individuals from the LGBTQ+ community.
10. **Endorsing white nationalism / white supremacy**: posts expressing support or sympathy with the white nationalist movement, its ideology, and activities. Includes comments related to the theme "make (white) America great again."
11. **Guns**: posts mentioning guns, weapons, or gun violence, as well as gun control and gun rights.

Table A1: Out of Sample Performance: Naive Bayes

	Accuracy	Precision	Recall	F1
Hateful language	0.94	0.98	0.96	0.97
Anti-Black	0.98	1.00	0.98	0.99
Anti-Jewish	0.98	0.99	0.99	0.99
Anti-Muslim	0.99	0.99	0.99	0.99
Anti-Asian	0.99	1.00	0.99	0.99
Anti-Latino	1.00	1.00	1.00	1.00
Anti-Immigrant	0.99	1.00	0.99	0.99
Misogyny	0.99	1.00	0.99	0.99
Anti-LGBTQ+	0.98	0.99	0.99	0.99
Pro white supremacy	0.98	1.00	0.99	0.99
Guns	0.98	0.99	0.99	0.99

Table A2: Summary statistics (Gab full sample)

	N	Mean	St. Dev.	Min	Max
Hate speech index	61,985,524	3.211	2.134	0.000	9.932
Hateful language	61,985,524	0.420	0.259	0.000	1.000
Anti-Black	61,985,524	0.339	0.262	0.000	1.000
Anti-Jewish	61,985,524	0.327	0.263	0.000	1.000
Anti-Asian	61,985,524	0.295	0.252	0.000	1.000
Anti-Muslim	61,985,524	0.307	0.258	0.000	1.000
Anti-Latinx	61,985,524	0.268	0.256	0.000	1.000
Anti-Immigrant	61,985,524	0.305	0.257	0.000	1.000
Misogyny	61,985,524	0.307	0.262	0.000	1.000
Anti-LGBTQ+	61,985,524	0.313	0.262	0.000	1.000
Pro-white supremacy	61,985,524	0.330	0.256	0.000	1.000
Guns	61,985,524	0.312	0.251	0.000	1.000

Figure A1: Posts Mentioning Facebook and YouTube on Gab

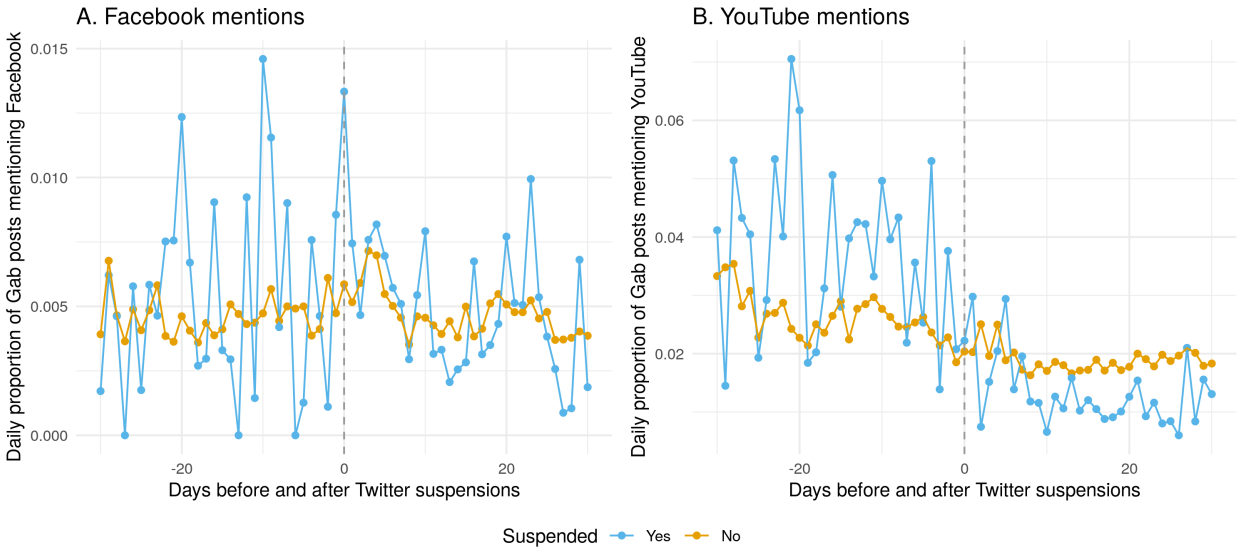


Table A3: Examples of Gab Users Talking About Big Tech Bias

1	*Republicans move to revoke Big Tech's 'free pass to censor and silence' conservatives*
2	*You can't hide your bias anymore, Google!*
3	*Looking like anti-maskers and anti lockdown supporters are to be classed as extremists. This hate from silicon valley big tech is despicable.*
4	*Watch Democrat Pulls The Most EVIL Sh* I have Ever Seen, Calls On Tech Monopoly To RESTRICT Our RIghts on YouTube*
5	*Facebook refugee. I've been incresingly disgusted with the left leaning bias found in the FB moderators and with FB's continual censorship of conservative thought. So, here I am on GAB. I'm looking forward to learning the program and making some new friends*
6	*White House drafting executive order to tackle Silicon Valley's alleged anti-conservative bias*