PNAS Nexus, 2023, 2, 1–10
https://doi.org/10.1093/pnasnexus/pgad346
Advance access publication 25 October 2023
Research Report

The systemic impact of deplatforming on social media

Amin Mekacher[a,1], Max Falkenberg[a,*,1] and Andrea Baronchelli[a,b,*]

[a]Department of Mathematics, City University of London, London EC1V 0HB, UK
[b]The Alan Turing Institute, British Library, London NW1 2DB, UK
*To whom correspondence should be addressed: Emails: max.falkenberg@city.ac.uk (M.F.); andrea.baronchelli.1@city.ac.uk (A.B.)
[1]A.M. and M.F. contributed equally to this work.
Edited By: David Rand

Abstract

Deplatforming, or banning malicious accounts from social media, is a key tool for moderating online harms. However, the consequences of deplatforming for the wider social media ecosystem have been largely overlooked so far, due to the difficulty of tracking banned users. Here, we address this gap by studying the ban-induced platform migration from Twitter to Gettr. With a matched dataset of 15M Gettr posts and 12M Twitter tweets, we show that users active on both platforms post similar content as users active on Gettr but banned from Twitter, but the latter have higher retention and are 5 times more active. Our results suggest that increased Gettr use is not associated with a substantial increase in user toxicity over time. In fact, we reveal that matched users are more toxic on Twitter, where they can engage in abusive cross-ideological interactions, than Gettr. Our analysis shows that the matched cohort are ideologically aligned with the far-right, and that the ability to interact with political opponents may be part of Twitter's appeal to these users. Finally, we identify structural changes in the Gettr network preceding the 2023 Brasília insurrections, highlighting the risks that poorly regulated social media platforms may pose to democratic life.

Significance Statement

Deplatforming, or banning harmful users from social media, is a common way to maintain online safety. Here, we study what happens to banned Twitter users who move to Gettr, a fringe platform popular with the far-right, finding that they post similar content but are more active than users who have not been banned. Increased Gettr use is not associated with increased toxicity. Rather, users with access to both platforms tend to be more toxic, and up to 7 times more active, on Twitter, where they retain the ability to engage with political opponents. The study highlights the impact of deplatforming on user behavior and raises questions about how best to regulate social media as an interconnected ecosystem.

Introduction

Social media has always been controversial, with constant debate around which content should be permitted, which content should be banned, and the conditions under which a user should be deplatformed for breaking the rules (1, 2). Particularly since the US Capitol insurrections, the deplatforming question has become a cornerstone of the polarized public discourse, with major social media companies facing increased pressure to deplatform malicious users (3, 4).

The rationale behind deplatforming is straightforward: Removing malicious accounts from social media helps protect other users and limits the spread of content which has the potential to cause harm (5–7). The scientific literature supports this view showing that many harmful communities are no longer active on mainstream platforms; these groups previously thrived by posting hate speech or conspiracy theories (5, 7–18), with their dense interaction networks facilitating a broad reach for their content (19).

However, the benefits of deplatforming for users on mainstream platforms do not account for the impact of these moderation policies on the wider social media ecosystem. For instance, a previous analysis showed that users banned from Twitter and Reddit became more active, and produced more toxic content, after joining the fringe platform Gab (7). Such fringe platforms are weakly regulated and poorly monitored, with evidence to suggest that they allow violent narratives to develop and thrive (10, 20–25). Despite this, the extent to which banning accounts from one platform drives migrations towards fringe alternatives remains unclear, and there has been no like-for-like comparison between banned users and politically aligned users who have not been suspended from the mainstream. This is in part because data is rarely available which permits the cross-platform tracking of social media users.

In this paper, we present a unique dataset which addresses this gap, focusing on a matched cohort of users who migrated from Twitter to Gettr, a Twitter-clone that has attracted many of Twitter's most high-profile suspended accounts including US congresswoman Marjorie Taylor-Greene, media executive Steve Bannon, and conspiracy theorist Alex Jones.

Our dataset presents the near-complete evolution of Gettr from its founding in July 2021 to May 2022 including 15M posts from

Competing Interest: The authors declare no competing interest.
Received: June 5, 2023. **Accepted:** October 6, 2023

785,360 active users who have posted at least once. Of these users, 6,152 are verified, 1,588 of which self-declare as active on Twitter (see Materials and methods section). For these 1,588 self-declared Twitter users with a verified Gettr account, we download their Twitter timeline from July 2021 to May 2022 totaling 12M tweets and retweets. These users represent the "matched" cohort, with analysis of their Gettr posts (Twitter tweets) referred to as "matched Gettr" ("matched Twitter") below. A manual check of these accounts confirms that 95% of matches across the two platforms are accurate, corresponding to the same individual or organization (see Supplementary Material). For the remaining verified Gettr users, we use the Twitter API to identify those accounts which have been suspended from the platform, assuming accounts share the same username on both platforms, totaling 454 accounts who constitute the "banned" cohort. Finally, all remaining users who are not verified on Gettr are part of the "nonverified" cohort.

In the remainder of the paper, we will overview account activity and retention on Gettr, showing that the banned cohort are 5 times more active than the matched cohort. Despite this, our results will show that these two cohorts are structurally mixed on Gettr, sharing the same politically homogenous audience and posting similar content. Using matched cohort tweets, we will show that Gettr is generally representative of the US far-right, and that matched users are more toxic on Twitter than they are on Gettr. We find little evidence to support the view that users become more toxic as a result of their extended use of the fringe platform. Finally, we will highlight how Gettr had a global impact, outlining the structural changes in the Portuguese-language Gettr network that emerged in the run up to the January 2023 riots in Brazil.

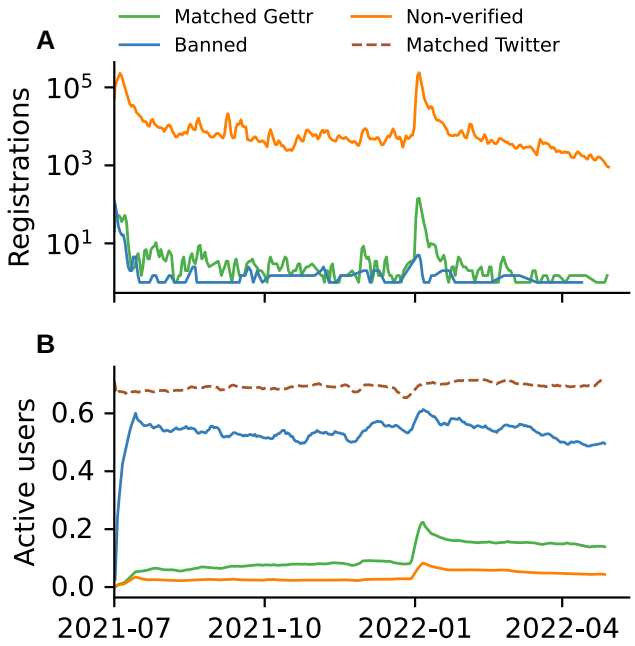

Fig. 1. User registrations and daily activity for each cohort. A) Three-day moving average of the daily number of users who registered on Gettr. The curve is displayed separately for the banned cohort, the matched cohort and other nonverified users on Gettr. B) Seven-day moving average of the proportion of users from each cohort who were active on Gettr on a given day. The percentage of the matched cohort active on Twitter is also shown (dashed).

Results

User acquisition and activity

We start by analyzing how the three cohorts of "matched Gettr," "banned," and "nonverified" users joined Gettr. Figure 1A shows that user registrations were largely steady over time with two exceptions where registrations peaked: (1) July 2021 when the platform was founded and (2) January 2022 after the suspension of Marjorie Taylor-Greene and Robert Malone on Twitter (26), and following the announcement by Joe Rogan that he would be opening a Gettr account (27).

In Fig. 1B, we show the fraction of accounts from each cohort who are active on any given day. For the matched cohort, we present their activity on both Gettr and Twitter. Focusing on the nonverified cohort, we see that a growing user base does not correlate with the growth of an engaged community, with, on average, 4% of the nonverified cohort active on any given day. On Gettr, 10% of the matched cohort are active on average, likely exceeding the value for the nonverified cohort because verified social media users are typically more active than other users (28). However, on Twitter the matched cohort are significantly more active with 69% of accounts active any given day. The activity of the matched cohort on Twitter is stable, with no evidence of a reduction in activity following the January 2022 suspensions. For the banned cohort on Gettr, activity approaches the baseline of the matched cohort on Twitter, with 53% active daily, 5 times larger than for the matched cohort on Gettr, and 13 times larger than the nonverified cohort. These results are qualitatively robust if we consider exclusively English-language accounts or Portuguese-language accounts, the first and second largest Gettr demographics, respectively (see Supplementary Material). A previous study has also shown a similar increase in activity for banned Twitter and Reddit users active on Gab (7).

User retention on Gettr

We now focus on the retention of users on Gettr. In Fig. 2, panels A and B show the survival curves for the proportion of users who remain active a certain number of days after registration (see Materials and methods section) for key registration months (July 2021 and January 2022 where registrations peaked, see Fig. 1), while panel C shows the average retention of users in each cohort over time. Survival curves for other registration months are shown in the Supplementary Material and follow the same pattern with higher banned retention than matched retention. The matched cohort are consistently active on Twitter with no evidence that users stop using the platform over time: 90% of the matched cohort are active in the first month covered by our dataset (July 2021), and 98% of these users remain active in our dataset's final month (April 2022). This highlights that the matched cohort are established Twitter users who are committed to the platform.

Figure 2 shows that the banned cohort have the highest retention on Gettr, independent of the month in which they joined the platform, whereas the nonverified cohort and the matched cohort become inactive at a faster rate. For the highlighted registration months, we note that the January curves fall off at a sharper rate than the July curves: For the July cohort, half of the newly registered users from the nonverified cohort become idle after 216 days, compared to only 68 days for the January cohort.

The event which clarifies these differences is the Marjorie Taylor-Greene deplatforming on Twitter. This deplatforming was denounced by Joe Rogan who opened a Gettr account on 2022 January 2, resulting in a large migration of his supporters, and supporters of Marjorie Taylor-Greene, to Gettr. However, after

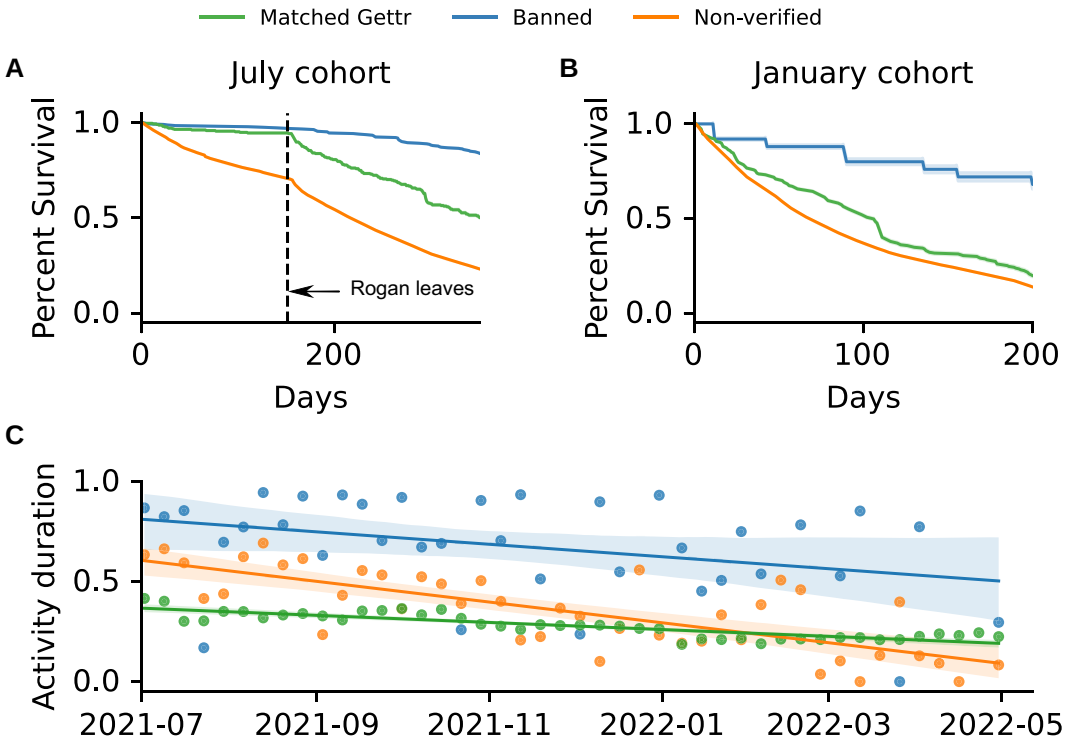

Fig. 2. User retention for key registration months and average retention by registration date over time. A) Kaplan–Meier survival curves for each user cohort showing the fraction of accounts who registered in July 2021 who remain active on Gettr a given number of days after registration for the banned cohort, matched cohort, and the nonverified cohort. The standard error of each curve is computed using Greenwood's formula (29) (see Materials and methods section). The dashed line corresponds to 2022 January 1, shortly before Joe Rogan joined Gettr. B) Survival curves for January 2022. C) Decay curves for user activity, showing the duration of their activity with respect to their registration date, normalized by the number of weeks to the end of our data collection period. Data for each cohort is fitted using linear regression ($y = ax + b$, $a = -0.007$, $[-0.014, 0]$, $b = 0.8$, $[0.65, 0.95]$ for banned users, $a = -0.011$, $[-0.015, -0.008]$, $b = 0.6$, $[0.52, 0.67]$ for matched users, and $a = -0.003$, $[-0.004, -0.002]$, $b = 0.36$, $[0.34, 0.37]$ for nonverified users; square brackets indicate 95% confidence interval, highlighted by shaded area).

criticizing the platform's policies (30), Rogan quit the platform on January 12. This 10-day period highlights how a single celebrity's endorsement resulted in a large migration to Gettr. However, the subsequent denouncement by Rogan not only resulted in many new users quitting the platform (those from the January 2022 cohort in panel B), but also resulted in many existing users quitting, see dashed line in panel A. Importantly, members of the banned cohort who registered in July 2021 did not leave Gettr at an enhanced rate after January 2022. This highlights that users who had the option to return to Twitter did so, but those who could not (due to suspension) continued to use Gettr.

Compared to previous Gettr studies which showed that users become idle shortly after registration (31), possibly due to the lack of engaging content (32), our results reveal the discrepancy between users banned from Twitter and users who remain active on Twitter, indicating that Gettr was most successful at retaining users who had lost their Twitter audience. Our results also show that deplatforming events of exceptional prominence can induce a significant influx of accounts into a fringe platform, but not necessarily a corresponding outflux from the dominant mainstream platform.

Gettr structure and content

In order to further clarify differences between banned and matched users, we now focus on the structure and content of the Gettr social network. We start by generating a topic model using Gettr posts (33) (see Materials and methods section). A table of topics and their description is provided in the Supplementary Material. This shows that content on Gettr is dominated by issues of broad relevance to the US political right including (1) Covid-19—one-sixth of all Gettr content, approaching one-third in some months—(2) deplatforming from Twitter and other social media platforms, (3) accusations of election fraud and the January 6th insurrection, and (4) broader issues regarding gender, abortion, gun-control, the US Supreme Court, and race.

Most topics discussed on Gettr are prominent in tweets authored by the matched cohort, however, three themes are disproportionately prominent on Gettr: (1) accusations of election fraud surrounding the 2020 US election, (2) resistance to Covid-19 vaccine mandates, particularly in relation to the "Freedom Convoy" protests in Canada, and (3) the Russian invasion of Ukraine. These are topics which are known to have been targets of the Twitter content moderation team (34–36).

We now measure whether the banned and matched cohorts are structurally segregated (or polarized) to assess whether the cohorts share the same, or different, audiences on Gettr. This check is important since there is no a priori reason to assume that the banned and matched cohorts are drawn from the same ideological group. We measure structural segregation (or polarization) using the latent ideology, a well-established method which constructs a synthetic ideological spectrum from user interactions on the platform (37–39) (see Materials and methods section). This measure orders the network of interactions between a set of influencer accounts (the banned and matched cohorts combined) and a set of accounts who interact with them (the nonverified

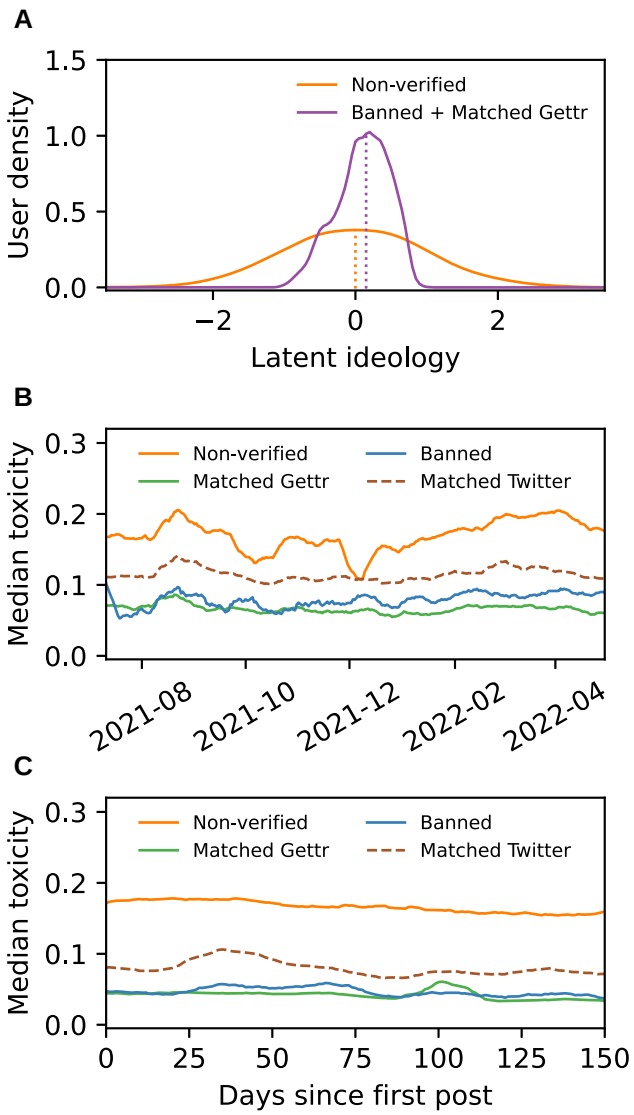

Fig. 3. The latent ideology of Gettr users, and the toxicity of Gettr posts and Twitter tweets. A) The latent ideology is calculated using the 500 most active banned and matched users on Gettr, merged into a single influencer cohort. Unit values on the x-axis correspond to the standard deviation of the ideology distribution for all users. Both distributions are unimodal when tested using Hartigan's diptest (multimodality not statistically significant for the nonverified cohort, $P = 0.99 > 0.01$, or banned and matched cohort, $P = 0.61 > 0.01$). B) The median post toxicity each day for each user cohort (14-day moving average). Toxicity is calculated using the Google Perspective API (40) (see Materials and methods section). Median toxicity [lower and upper quartile] for the nonverified cohort, 0.17 [0.06, 0.37], banned cohort, 0.05 [0.02, 0.15], matched cohort on Gettr, 0.04 [0.02, 0.11], and matched cohort on Twitter, 0.09 [0.04, 0.22]. C) The median toxicity of posts authored a fixed number of days after a user first posted on Gettr (or Twitter; 14-day moving average). There is minimal evidence of a meaningful increase in user toxicity due to extended Gettr use (see Supplementary Material).

cohort). By merging the banned and matched cohort into a single group, we can measure differences in how the nonverified cohort interact with banned and matched users in an unbiased manner based on purely structural factors. Note that we exclude a small number of non-US- based accounts from the influencer set to avoid geographical conflation (see Materials and methods section).

The distribution of the latent ideology for the banned and matched cohort, and for the nonverified cohort, is shown in Fig. 3A. Both distributions are unimodal according to Hartigan's diptest (41). We observe that the banned and matched distribution falls within the bounds of the broader nonverified distribution. The banned and matched distribution is, however, significantly narrower, a feature indicative of the network centrality of these users who play a central role in the general Gettr discussion. Nonverified Gettr users are found both at the core of the Gettr discussion and at the peripheries. The central role of banned and matched users is expected since verified social media accounts typically attain higher engagement than nonverified accounts (42, 43).

The unimodal ideology, and the central position of the matched and banned cohorts, indicates that these users share a common audience on Gettr; segregated audiences would appear as a multimodal ideology distribution [see examples in Refs. (38, 39)].

Content toxicity and Twitter mentions

We now focus on the toxicity of posts from each user cohort, shown in Fig. 3B. Toxicity is calculated using the Google Perspective API (40) (see Materials and methods section). The panel shows the median daily toxicity of each cohort. By comparing the toxicity distributions using a Kolmogorov-Smirnov (KS) test, and by applying a bootstrapping procedure to ensure equal sample sizes (see Materials and methods section), we find that posts authored by the nonverified cohort are more toxic than posts by the matched cohort (KS-test $D = 0.36$, $P = 1.3 \times 10^{-57}$), and than posts by the banned cohort (KS-test $D = 0.32$, $P = 8.0 \times 10^{-47}$). We also find that the tweets authored by the matched cohort are more toxic than Gettr posts authored by the matched cohort (KS-test $D = 0.23$, $P = 8.7 \times 10^{-23}$). However, the difference between the toxicity of posts for the banned cohort and matched cohort on Gettr is not statistically significant (KS-test $D = 0.06$, $P = 0.09$).

In order to assess whether individual users are becoming more toxic over time due to their extended Gettr use, we plot the median toxicity of posts authored a fixed number of days after each user first posted on Gettr (or first posted on Twitter during our observation period), as shown in Fig. 3C. For the nonverified cohort, the gradient in the change of the toxiciy over time is not significantly different from zero (see Supplementary Material). For the other cohorts, there is a statistically significant but very small nonzero gradient in the toxicity over time. However, this gradient is negligible when considered in the context of the much larger inter-quartile range of post toxicity values for each cohort (see Supplementary Material).

Together, the results for the latent ideology, topic modeling, and toxicity show that, although there are significant differences in activity and retention between the banned and matched cohorts on Gettr (see Figs. 1 and 2), there is little that distinguishes their audience and content. This result confirms previous research which shows that fringe platforms are politically homogeneous; platforms with this property may be referred to as "echo-platforms" (44, 45). In contrast, mainstream platforms are often politically diverse, but with opposed political groups confined to echo-chambers (39, 43, 45–49).

Considering the toxicity of posts for each topic, we find that topics with disproportionately high toxicity are related to race (e.g. Black Lives Matter; median post toxicity [lower and upper quartile] =0.40 [0.31, 0.52]), focus on female US Democratic politicians (0.38 [0.18, 0.58], and discuss gender issues (0.38 [0.24, 0.51]). All three topics are known to attract abusive content on social media (50–52).

We now explore possible reasons why the matched cohort are more toxic on Twitter than they are on Gettr. To do this, we

analyze the Twitter accounts mentioned in tweets authored by the matched cohort. For each mentioned account, we compute the ratio between the number of users from the matched cohort who quote-tweet that account and the number of users from the matched cohort who quote-tweet or retweet that account. This ratio (referred to as the "quote-ratio" throughout) is instructive since there is evidence that retweets are often (but not exclusively, journalists being a known exception) used to endorse the message of the original author (39, 53), whereas quote tweets allow a user to comment on a message in either a positive, negative, or neutral manner. Negative "quoting" behavior is a known method of communication with ideological opponents across polarized environments (54, 55). Hence, a low quote-ratio (i.e. the account is disproportionately retweeted) indicates general endorsement by the matched cohort of users, whereas a high quote-ratio (i.e. the account is disproportionately quote-tweeted) indicates that the matched cohort are more likely to disagree with and hold a negative view of this account.

Figure 4 shows the toxicity of tweets authored by the matched cohort, binned according to their quote-ratio. We count each mentioner-mentionee pair only once for quote-tweets and once for retweets to avoid bias from highly active accounts, and only include accounts mentioned by at least five matched users. This reveals (1) that tweets authored by the matched cohort mentioning any Twitter account are more toxic than tweets which do not mention another account and (2) that tweets authored by the matched cohort are more toxic if they mention an account with a high quote-ratio than if they mention accounts with lower quote-ratios.

To better understand this result, we plot the distribution of the quote-ratio broken down into four groups. Figure 5A shows the distribution of all users mentioned by the matched cohort, and the distribution for Twitter accounts who are also part of the matched cohort (i.e. a matched account mentioning another matched account). Three individuals are marked on the figure: (1) Republican 2022 Senate nominee Herschel Walker, the user with the lowest quote-ratio of prominent mentioned accounts (>100 unique mentions), (2) Democratic speaker of the house Nancy Pelosi,[a] the user with the largest quote-ratio (>100 unique mentions), and (3) Elon Musk, the account with the most unique mentions.

Figure 5B shows elected US political accounts mentioned by the matched cohort, labeled using the dataset in (56), broken down by party affiliation. This shows that Republican politicians are disproportionately retweeted (i.e. endorsed) by the matched cohort, whereas Democrats are disproportionately quote-tweeted. The individuals marked on this panel are political outliers; Liz Cheney and Adam Kinzinger are the Republican politicians with the highest quote-ratios (>10 unique mentions), whereas Tulsi Gabbard[b] and Kyrsten Sinema are the Democratic politicians with the lowest quote-ratios (>10 unique mentions). This shows that these politicians do not align with the dominant position of their parties. Consequently, the matched cohort are more likely to endorse the Democratic outliers, and more likely to negatively quote-tweet the Republican outliers; Liz Cheney and Adam Kinzinger have been referred to as RINOs ("Republicans in name only") by their far-right opponents (57, 58).

Figure 5C shows the news media organizations mentioned by the matched cohort, grouping them according to their political leaning as classified by Media Bias/Fact Check (MBFC; see Materials and methods section). Previous research confirmed that MBFC classifications are similar to classifications from other reputable media rating organizations (59). Finally, Fig. 5D repeats the analysis in panel C, but groups media outlets according to whether MBFC labels them as reliable or questionable.

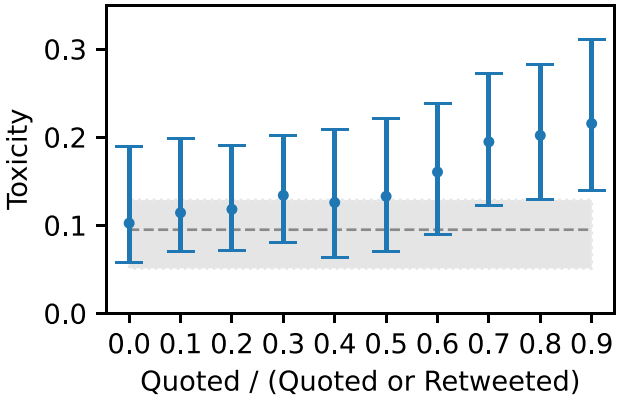

Fig. 4. Toxicity of tweets authored by the matched cohort mentioning other Twitter accounts, binned according to their quote-ratio. The distribution of the quote-ratio is shown in Fig. 5. Each point indicates the median toxicity of tweets with a quote-ratio within the binned range $[x, x+0.1)$. Error bars indicate the inter-quartile range. The dashed line indicates the median toxicity for all tweets (including those which do not mention another account) from the matched cohort, with the shaded region indicating the inter-quartile range; all data points lie above this line.

Using the distribution of all mentions (the "any user" curve in Fig. 5A) as the baseline behavior of the matched cohort, we find that, when tested using a two-sample Kolmogorov–Smirnov test, only the distributions of far-right media organizations in panel C (KS-test P–value $= 0.24 > 0.01$; Cohen's $d = 0.20$) and questionable media organizations in panel D (KS-test P–value $= 0.29 > 0.01$; Cohen's $d = 0.05$) are not significantly different from the baseline (see Supplementary Material). This observation aligns with previous research showing that the US political right on Twitter are more likely to share questionable news sources, and are more likely to be suspended (60).

The Democrat politicians distribution has the largest statistical difference to the all-mention baseline (KS-test P–value $= 3 \times 10^{-16} < 0.01$; Cohen's $d = 2.34$). With the exception of Tulsi Gabbard, no Democratic politicians has a known Gettr account; 132 are mentioned on Twitter by the matched Gettr cohort. In contrast, 32 Republican politicians have been active on Gettr; 151 are mentioned on Twitter by the matched Gettr cohort.

Combining the evidence from the topic modeling and from the quote-ratio in Fig. 5 indicates that the matched cohort are aligned with the US far-right, often quote-tweeting, but not retweeting, their Democratic political opponents and moderate Republicans. In conjunction with the latent ideology in Fig. 3, this suggests that Gettr as a whole is generally representative of the US far-right. These results suggest that the ability to mention one's political opponents on Twitter is part of the reason that the matched cohort are more toxic on Twitter than they are on Gettr where direct interactions with political opponents are not possible (61, 62).

Discussion and conclusion

In this paper, we have analyzed self-declared user-level matching to study the ban-induced platform migration from Twitter to

[a] Speaker of the house during the timeframe of our dataset.
[b] Democrat during our analysis timeframe; left the Democratic party in October 2022.

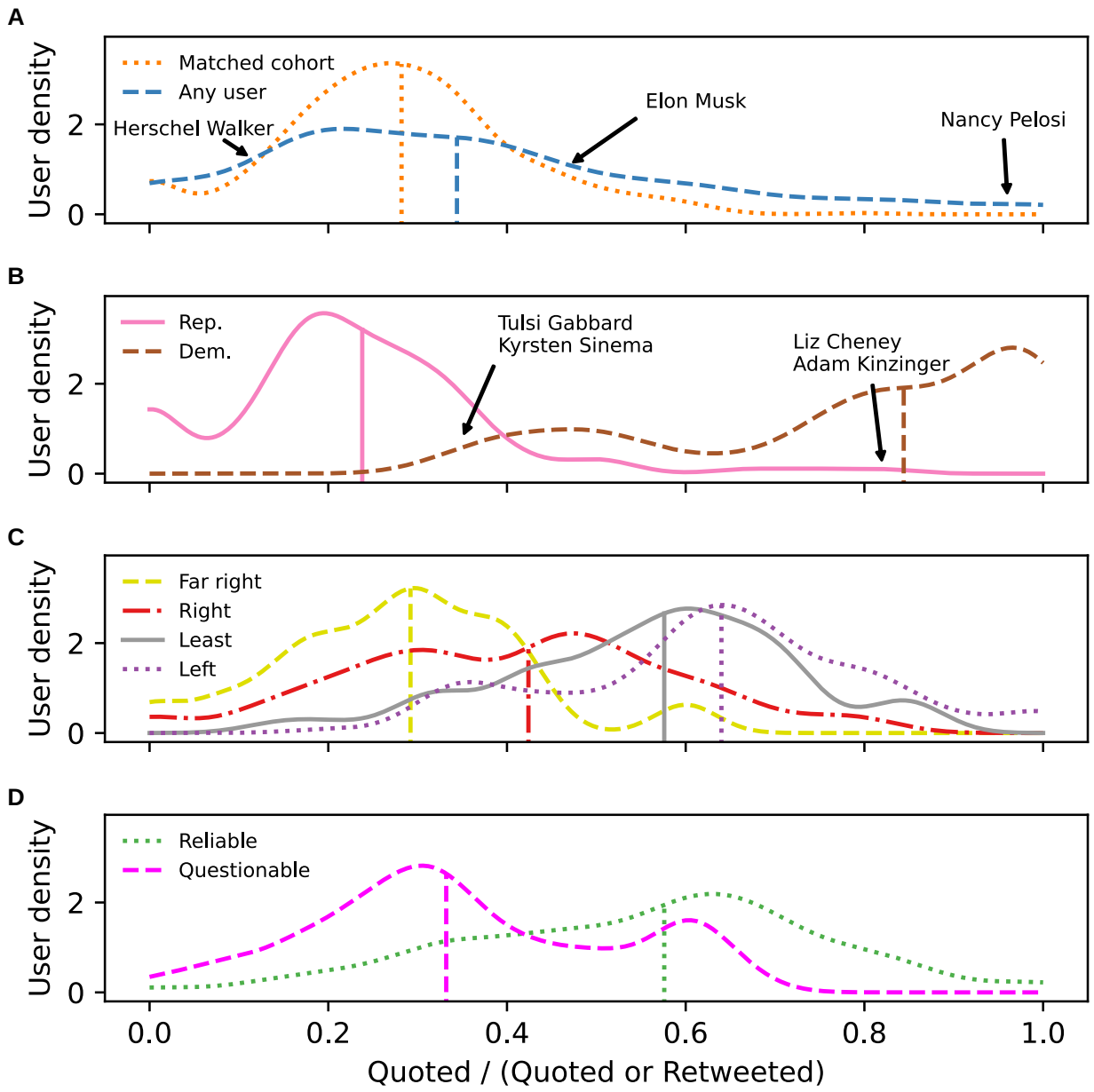

Fig. 5. The distribution of the quote-ratio of accounts mentioned on Twitter by the matched cohort. A) The quote-ratio distribution for all mentioned accounts (dashed), and for mentioned accounts who are part of the matched cohort of users (i.e. a matched user mentioning another matched user; dotted). B) The quote-ratio distribution for Twitter accounts belonging to known elected US Republican (solid) and known elected US Democrat (dashed) politicians. C) The quote-ratio distribution for Twitter accounts belonging to news media organizations who have been labeled with a political leaning by MBFC. Organizations are classified as left (dotted), least-biased (solid), right (dot-dashed), or far-right (dashed). D) The same news media organizations, but broken down according to whether they are classified as a reliable (dotted) or questionable (dashed) by MBFC. Vertical lines mark the median of each distribution. Annotations indicate mentioned accounts of particular interest (see text).

Gettr. First, we showed that the banned cohort of users deplatformed from Twitter are more active on Gettr, and have higher platform retention, than the matched cohort who remain active on Twitter. Second, we revealed that Gettr content primarily discusses themes relevant to the US political right. Topics overepresented on Gettr are known to have resulted in account suspensions on Twitter. Third, we showed that the matched and banned cohorts share the same politically homogeneous Gettr audience. Finally, we found that matched users are more toxic on Twitter than they are on Gettr, and that these toxic tweets often directly mention political opponents. We find little evidence of a meaningful increase in user toxicity over time. Finally, we highlighted Gettr's broader societal impact, revealing significant structural changes in the Gettr interaction network in the run-up to the Brasília insurrections.

The fact that the banned and matched cohorts appear similar in every regard, apart from their activity and retention on Gettr, is evidence of the systemic impact of deplatforming. Fringe platforms offer a safe haven where deplatformed users are free to capitalize on their supporters following suspension from Twitter. However, in this politically homogeneous environment, users are essentially confined to an ideological "echo-platform" (44, 45) where they cannot engage and confront their political opponents. Our results hint that this ability to interact with opponents may be part of Twitter's appeal for far-right social media users, although more work is needed to fully clarify this observation.

When users are banned from mainstream platforms, they become wholly dependent on the fringe alternatives [despite Gettr also suspending some users, notably white supremacist Nicholas Fuentes (63)]. This may pose a societal risk since fringe platforms are believed to facilitate the emergence of radical narratives and the spread of hate speech (64, 65). A lack of monitoring can, therefore, mean that signs of collective upheaval are missed. The Brasília insurrection, which took place in January 2023 following Jair Bolsonaro's defeat in the presdential election, is an example of this, as the election fraud allegations were widely spread by Steven Bannon's podcast *The War Room*, streaming regularly on Gettr while being banned from Twitter (66, 67); an analysis provided in the Supplementary Material shows how Portuguese-language Gettr activity rose in the weeks preceding the riots.

These results complement and add to existing work which considers the effect of mainstream deplatforming on users' behavior on Gab primarily on the basis of content analysis (7). However, while the Gab study finds an increase in user toxicity over time, we find little evidence of a similar increase on Gettr. This difference may indicate that changes in toxicity (or lack thereof) depend on the fringe platform used after suspension, rather than on deplatforming itself, but more work is needed to validate this hypothesis in the context of the wider social media ecosystem.

It is important to contextualize the scope of our findings, whose limitations are avenues for future work. First, the current study only considers the migration from Twitter to Gettr, since users of other platforms do not declare their Twitter use as standard. If data becomes available, future work should extend scrutiny to multiple other platforms, ideally in a unified study. Second, the Gettr matching feature only applies to verified users, a subset of the users who migrated from Twitter to Gettr. Analyzing nonverified users migrating across platforms would clarify the differential impact of deplatforming on content creators as opposed to consumers. Finally, we cannot study tweets from the banned cohort since this data is not publicly available. Analyzing this content would explain why some users are deplatformed, but others are not.

Overall, our study highlights how Gettr struggles to compete with Twitter when users have free choice to use either platform. However, the decision by Twitter to deplatform a user impacts how those users view Gettr as an alternative. We anticipate that future work will build on these observations and speculate that other fringe platforms will likely show a similar dependence on their mainstream competitors. This work is urgently needed given the potential risks posed to democracy by poorly regulated social media (68, 69).

Materials and methods

Gettr data

The data used for this study has been collected using GoGettr (70), a public client developed by the Stanford Internet Observatory to give access to the Gettr API. This API allows to query user interactions, including the posts they like or share. User profiles were initially collected through a snowball sampling, by using highly popular accounts on the platform as seed users, and using the API to query their follower and following list, before repeating the same process for a random sample of the newly retrieved users. Repeating this process many times ensures that our dataset is near-complete for the studied time period.

To attract more users from Twitter, Gettr previously offered a feature that would automatically import a user's tweets upon creating an account. However, due to Twitter blocking this capability on 2021 July 10 (71), Gettr had to discontinue this feature. To ensure the accuracy of our results, any posts imported before 2021 July 10, and any Gettr post whose timestamp precedes the account's creation date were removed from our dataset.

To ensure our case study on the Brazilian right encompasses the Brasília insurrection, we expanded the data collection time frame for any user associated with the Brazilian community. The data collection was run in July 2022 for every user whose profile we have retrieved, and in January 2023 for the users in the Brazilian cohort.

Twitter data

For each verified Gettr account where the Gettr API references their Twitter followers in the account metadata, we check that the Twitter account with the same username is active using the Twitter API (see https://developer.twitter.com/en/products/twitter-api/academic-research). We identify accounts who were previously active on Twitter but are now banned from the "HTTP 403 Code 63" error message corresponding to suspended Twitter accounts. Other error messages are used for protected or not found accounts. Our study does not consider users banned from Twitter who did not join Gettr, or joined Gettr using a different username to their original Twitter username.

For each active account we download their Twitter timeline including all tweets and retweets in the period July 2021 to May 2022. This totals approximately 12 million tweets. Data was collected between September and October 2022, preceding Elon Musk's amnesty of suspended Twitter accounts.

User labeling

Throughout our analysis, we label verified Gettr users as being either "matched" or "banned," depending on whether their corresponding Twitter account is active or suspended. Any verified user who decided to link their Twitter account on their Gettr page has their Twitter follower count displayed on their profile, which can also be retrieved from the Gettr API. We stress that this self-declaration permits cross-platform matching since users can "reasonably expect" that their Gettr accounts will be associated with their Twitter accounts.

To match accounts across platforms, we assumed that users picked the same username on both Gettr and Twitter, and we used the Twitter API to retrieve their Twitter activity. A user is classified as "banned" if the Twitter user endpoint returned an error indicating that the account has been suspended (Error "HTTP 403 Code 63").

To validate the accuracy of this matching process, we manually check whether matched accounts correspond to the same organization or individual on both Gettr and Twitter (see Supplementary Material). This reveals a 95% matching accuracy. Note that following Elon Musk's amnesty on banned accounts, approximately one-third of accounts in the banned cohort have been reactivated. Of the top 100 most followed accounts on Gettr from the banned cohort, 33 have been unbanned on Twitter following the Musk amnesty, and all 33 are an exact match for the same individual or organization across the two platforms. However, due to the discontinuation of the Twitter API for academic purposes the detailed analysis of these unbanned accounts on Twitter is not possible.

For data privacy reasons, all analysis of users across platforms is aggregated at the cohort level; we do not present results for individual users.

Calculating account survival

The Kaplan–Meier estimate is a tool used to quantify the survival rate of a population (in our case, users active on a social platform) over time. For each time step t, we measure how many users become indefinitely inactive, and we quantify the survival rate as

$$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right), \quad (1)$$

where d_i represents the number of users who became inactive at time t_i and n_i is the number of users who are still active up to time t_i. Greenwood's formula is used to estimate the confidence interval for the Kaplan–Meier estimate of the survivor function. For the study time t, the standard error is given by

$$\widehat{\mathrm{SE}}_2(t) = \sqrt{\hat{S}^2(t) \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}}. \quad (2)$$

Topic modeling

Gettr and Twitter content is analyzed using the BERTopic topic modeling library (33). This method extracts latent topics from an ensemble of documents (in our case Gettr posts and Twitter tweets). The base model uses pretrained transformer-based language models to construct document embeddings which are then clustered.

These methods are known to struggle with very short documents which are common on microblogging sites. Hence, we train our topic model using exclusively Gettr posts which are longer than 100 characters. To avoid any single user dominating a specific topic, we limit the training set to no more than 50 posts from any given user.

Latent ideology

To calculate the latent ideology on Gettr, we use the method developed in Refs. (37, 38) and filtering procedures from Ref. (39). The method uses a bipartite approach where it classifies Gettr accounts as influencers or regular users. It then generates an ordering of users based on the interaction patterns of regular users with the influencer set. In the current paper, we select the matched and banned user sets as our influencers, and the remaining set of Gettr users as our cohort of regular users.

Two factors can conflate the latent ideology: (1) account geography and (2) a lack of user-influencer interactions. Since we are interested in the segregation of the banned and matched cohorts based on political ideology, the former is problematic because country-specific communities on social media can appear structurally segregated from a related community in other countries, even if they are politically aligned. For this reason, we remove a small number of accounts associated with the UK and China from our set of Gettr influencers. In the latter case, a lack of user-influencer interactions can be problematic since influencers with few user interactions appear as erroneous outliers when computing the latent ideology, often because they do not take active part in the conversation. Hence, we restrict our influencer set to the 500 banned and matched accounts who receive the largest number of interactions from other users on Gettr. In the current study, we consider any interaction type including comments, shares and likes. The latent ideology is robust as long as the number of influencers used is larger than 200 accounts (39).

In order to assess the modality of the ideology distributions, we use Hartigan's diptest. This approach is used to identify polarized social media conversations and echo-chambers (38, 39). Hartigan's diptest compares a test distribution against a matched unimodal distribution to assess distribution modality (41).

The test computes the distance, D, between the cumulative density of the test distribution and the cumulative density of the matched unimodal distribution. The D-statistic is accompanied by a P-value which quantifies whether the test distribution is significantly different to a matched unimodal distribution. A P-value of less than 0.01 indicates a multimodal distribution.

Toxicity analysis

The toxicity of Gettr and Twitter content is computed using the Google Perspective API (40), which has been used in several social media studies to assess platform toxicity (72–74).

Given a text input, the API returns a score between 0 and 1, indicating how likely a human moderator is to flag the text as being toxic. For our analysis, we used the flagship attribute "toxicity," which is defined as "[a] rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion" (75).

When computing statistics for the toxicity of each user cohort, we apply a bootstrapping procedure to avoid erroneous results from variable cohort sizes. This is important since the distribution of post toxicity is fat tailed; there are far more posts with low toxicity than high toxicity. Therefore, a smaller set of posts from a user cohort may have a lower median toxicity, purely due to sampling effects. To avoid this conflation, bootstrapping is employed where equally sized samples are drawn from each cohort (usually 100 posts), and the median toxicity is computed for each sample. Then, the sampling procedure is repeated 100 times to compute the median and inter-quartile range for the sampled post toxicity.

News media classification using Media Bias/Fact Check

For the quote-ratio analysis in Fig. 5, we identify the Twitter handles of news media outlets and classify their political leaning using Media Bias/Fact Check (MBFC; see https://mediabiasfactcheck.com/). Ratings provided by MBFC are largely similar to other reputable media rating datasets (59).

MBFC classify news outlets under seven leaning categories: extreme left, left, center-left, least (media considered unbiased), center-right, right, extreme right. To ensure that we have enough news media outlets to enable the quote-ratio analysis, we group these classifications into four larger groups: Left—(extreme left, left, center-left), Least—(least), Right—(center-right, right), Far-right—(extreme right). Note that we have chosen to use the terminology "far-right" instead of "extreme right" since the former is more common in the academic literature.

Acknowledgments

M.F. thanks Alessandro Galeazzi for providing the Media Bias/Fact Check data.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

M.F. and A.B. acknowledge support from the IRIS Infodemic Coalition (UK government, grant no. SCH-00001-3391).

Author Contributions

All authors designed the research. A.M. and M.F. acquired the data and performed the measurements. All authors analyzed the data, wrote the manuscript, discussed the results, and commented on the manuscript.

Preprints

This manuscript was posted as a preprint on 2023 March 20 at arxiv.org/abs/2303.11147.

Data Availability

Gettr and Twitter data used for this project is available in anonymized format on request at https://osf.io/dx2p8/.

References

1 Owono J. 2022. Banning content platforms is not a solution to hate speech on the internet, even when the platform is meta [accessed 2023 Mar 16]. https://www.justsecurity.org/82601/banning-content-platforms-is-not-a-solution-to-hate-speech-on-the-internet-even-when-the-platform-is-meta/

2 UNESCO, U. N. O. on Genocide Prevention, and the Responsibility to Protect. 2021. Addressing hate speech on social media: contemporary challenges [accessed 2023 Oct 31]. https://unesdoc.unesco.org/ark:/48223/pf0000379177

3 League A-D. 2021. ADL calls on twitter to suspend accounts supporting QAnon [accessed 2023 Mar 11]. https://www.adl.org/resources/blog/adl-calls-twitter-suspend-accounts-supporting-qanon

4 Horta Ribeiro M, Hosseinmardi H, West R, Watts DJ. 2023. Deplatforming did not decrease Parler users' activity on fringe social media. *PNAS Nexus*. 2:pgad035. doi:10.1093/pnasnexus/pgad035

5 Jhaver S, Boylston C, Yang D, Bruckman A. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proc ACM Hum-Comput Interact*. 5. 1–30. doi:10.1145/3479525

6 Johansson P, Enock F, Hale S, Vidgen B, Bereskin C. 2022. How can we combat online misinformation? A systematic overview of current interventions and their efficacy. *ArXiv Preprint*. https://arxiv.org/abs/2212.11864

7 Ali S, *et al.* 2021. Understanding the effect of deplatforming on social networks. In: 13th ACM Web Science Conference, 2021. p. 187–195. doi:10.1145/3447535.3462637

8 Horta Ribeiro M, *et al.* 2021. The evolution of the manosphere across the web. *Proc Int AAAI Conf Web Soc Media*. 15:196–207.

9 Ribeiro MH, *et al.* 2021. Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/incels. *Proc ACM Hum-Comput Interact*. 5:1–24. doi:10.1145/3476057

10 Rauchfleisch A, Kaiser J. 2021. Deplatforming the far-right: an analysis of youtube and BitChute. *SSRN Electron J*. 1–28. doi:10.2139/ssrn.3867818

11 Voskresenskii V. 2023. Migrating counterpublics: German far-right online groups on Russian social media. *Int J Commun*. 17:926–946.

12 Mekacher A, Papasavva A. 2022. "I can't keep it up." A dataset from the defunct Voat.co news aggregator. *Proc Int AAAI Conf Web Soc Media*. 16:1302–1311.

13 Russo G, Ribeiro MH, Casiraghi G, Verginer L. 2022. Understanding online migration decisions following the banning of radical communities. arXiv, arXiv:2212.04765, preprint: not peer reviewed.

14 Evkoski B, Pelicon A, Mozetič I, Ljubešić N, Novak PK. 2022. Retweet communities reveal the main sources of hate speech. *PLoS ONE*. 17:e0265602. doi:10.1371/journal.pone.0265602

15 Urman A, Katz S. 2020. What they do in the shadows: examining the far-right networks on Telegram. *Inf Commun Soc*. 25:904–923. doi:10.1080/1369118x.2020.1803946

16 Bryanov K, Vasina D, Pankova Y, Pakholkov V. 2022. The other side of Deplatforming: right-wing telegram in the wake of trump's Twitter Ouster. In: *Communications in computer and information science*. St Petersburg, Russia: Springer International Publishing. p. 417–428.

17 Chandrasekharan E, *et al.* 2017. You can't stay here. *Proc ACM Hum-Comput Interact*. 1:1–22. doi:10.1145/3134666

18 Zhang X, Wei Z, Du Q, Zhang Z. 2022. Social media moderation and content generation: evidence from user bans. *SSRN Electron J*. 1–69. doi:10.2139/ssrn.4089011

19 Ribeiro M, Calais P, Santos Y, Almeida V, Meira W. 2018. Characterizing and detecting hateful users on twitter. *Proc Int AAAI Conf Web Soc Media*. 12:676–679. doi:10.1609/icwsm.v12i1.15057

20 Winter C, *et al.* 2020. Online extremism: research trends in internet activism, radicalization, and counter-strategies. *Int J Confl Violence (IJCV)*. 14(2). 1–20. doi:10.4119/IJCV-3809

21 Bovet A, Grindrod P. 2022. Organization and evolution of the UK far-right network on Telegram. *Appl Netw Sci*. 7:76. doi:10.1007/s41109-022-00513-8

22 Tufekci Z. 2018. How social media took us from Tahrir Square to Donald Trump [accessed 2023 Jan 26]. https://www.technologyreview.com/2018/08/14/240325/how-social-media-took-us-from-tahrir-square-to-donald-trump/

23 Cinelli M, *et al.* 2021. Dynamics of online hate and misinformation. *Sci Rep*. 11. 22083. doi:10.1038/s41598-021-01487-w

24 Naffakh M. 2022. How pro-terrorism accounts are circumventing moderation on social media [accessed 26 Jan 2023]. https://observers.france24.com/en/middle-east/20221125-social-media-propaganda-islamic-state-terrorism

25 Romano A. 2021. What we still haven't learned from gamergate [accessed 2023 Jan 26]. https://www.vox.com/culture/2020/1/20/20808875/gamergate-lessons-cultural-impact-changes-harassment-laws

26 Johnson K. 2022. Twitter permanently bans U.S. representative Marjorie Taylor-Greene [accessed 2023 Mar 17]. https://www.reuters.com/world/us/twitter-permanently-bans-us-representative-marjorie-taylor-greene-2022-01-02/

27 McKay T. 2022. Joe Rogan joined Gettr 10 days ago and already thinks it sucks [accessed 2023 Mar 17]. https://gizmodo.com/joe-rogan-joined-gettr-10-days-ago-and-already-thinks-i-1848346452

28 DeVerna MR, Aiyappa R, Pacheco D, Bryden J, Menczer F. 2022. Identification and characterization of misinformation superspreaders on social media. *Arxiv Preprint*. 1–15. https://arxiv.org/abs/2207.09524

29 Cantor AB. 2001. Projecting the standard error of the Kaplan–Meier estimator. *Stat Med*. 20:2091–2097.

30 Friedman D, Breland A. 2022. Leaked messages show Gettr in crisis mode over Joe Rogan criticism [accessed 2023 Jan 20]. https://www.motherjones.com/politics/2022/01/leaked-messages-show-gettr-crisis-mode-joe-rogan-jason-miller-guo-wengui/

31 Thiel D, McCain M. 2021. Topologies and tribulations of Gettr: A month in the life of a new alt-network. Stanford Internet Observatory [accessed 2023 Oct 31]. https://fsi.stanford.edu/publication/topologies-and-tribulations-gettr-month-lifenew-alt-network-0

32 Schwemmer C. 2021. The limited influence of right-wing movements on social media user engagement. *Soc Med Soc*. 7:20563051211041650.

33 Grootendorst M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv, arXiv*:2203.05794, preprint: not peer reviewed.
34 Woo E. 2022. Twitter will stiffen moderation policies in response to the war in Ukraine. The New York Times. https://www.nytimes.com/2022/04/05/business/twitter-policy-ukraine.html
35 Twitter Transparency Team. 2022. COVID-19 misinformation transparency report [accessed 2023 Feb 7]. https://transparency.twitter.com/en/reports/covid19.html#2021-jul-dec
36 Conger K, Isaac M. 2021. Twitter permanently Bans Trump, capping online Revolt. The New York Times. https://www.nytimes.com/2021/01/08/technology/twitter-trump-suspended.html
37 Barberá P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Polit Anal*. 23:76–91.
38 Flamino J, *et al.* 2023. Political polarization of news media and influencers on Twitter in the 2016 and 2020 US presidential elections. *Nat Hum Behav*. 7(6):904–916. doi:10.1038/s41562-023-01550-8
39 Falkenberg M, *et al.* 2022. Growing polarization around climate change on social media. *Nat Clim Change*. 12:1114–1121.
40 Jigsaw. 2023. Google Perspective API [accessed 2023 Feb 3]. https://perspectiveapi.com/
41 Hartigan JA, Hartigan PM. 1985. The dip test of unimodality. *Ann Stat*. 13:70.
42 González-Bailón S, De Domenico M. 2021. Bots are less central than verified accounts during contentious political events. *Proc Natl Acad Sci*. 118:e2013443118.
43 Bovet A, Makse HA. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nat Commun*. 10. 7. doi:10.1038/s41467-018-07761-2
44 Cinelli M, *et al.* 2022. Conspiracy theories and social media platforms. *Curr Opin Psychol*. 47:101407.
45 Cinelli M, Morales GDF, Galeazzi A, Quattrociocchi W, Starnini M. 2021. The echo chamber effect on social media. *Proc Natl Acad Sci*. 118:e2023301118.
46 Jamieson KH, Cappella JN. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. USA: Oxford University Press.
47 Barberá P. 2020. *Social media and democracy: the state of the field, prospects for reform*. Vol. 34. New York: Cambridge University Press.
48 Conover M, *et al.* 2011. Political polarization on Twitter. *Proc Int AAAI Conf Web Soc Media*. 5:89–96.
49 Cota W, Ferreira SC, Pastor-Satorras R, Starnini M. 2019. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Sci*. 8:35.
50 Gallagher RJ, Reagan AJ, Danforth CM, Dodds PS. 2018. Divergent discourse between protests and counter-protests: #BlackLivesMatter and #AllLivesMatter. *PLoS ONE*. 13:e0195644.
51 Mamié R, Horta Ribeiro M, West R. 2021. Are anti-feminist communities gateways to the far right? Evidence from Reddit and YouTube. In: 13th ACM Web Science Conference 2021, WebSci '21. New York (NY): Association for Computing Machinery. p. 139–147.
52 Krook ML. 2018. Violence against women in politics: a rising global trend. *Politics Gend*. 14:673–675.
53 Metaxas P, *et al.* 2015. What do retweets Indicate? Results from user survey and meta-review of research. Oxford: AAAI.
54 Guerra PC, Nalon R, Assunçao R, Meira W. 2017. In: Eleventh International AAAI Conference on Web and Social Media. Montreal: AAAI.
55 Stella M, Ferrara E, De Domenico M. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proc Natl Acad Sci*. 115:12435–12440.
56 Van Vliet L, Törnberg P, Uitermark J. 2020. The Twitter parliamentarian database: analyzing Twitter politics across 26 countries. *PLoS ONE*. 15:e0237073.
57 Shepherd B, Axelrod T. 2022. How Liz Cheney went from rising Republican star to primary underdog after Jan. 6. [accessed 2023 Feb 21]. https://abcnews.go.com/Politics/liz-cheney-rising-republican-star-enters-primary-underdog/story?id=88415555
58 Fedor L, Edgecliffe-Johnson A. 2021. Adam Kinzinger: the 'Rino' leading the charge for a post-Trump GOP [accessed 2023 Feb 21]. https://www.ft.com/content/7984b8e0-7c8b-4527-95ec-4e46b6048499
59 Lin H. 2023. High level of correspondence across different news domain quality rating sets. *PNAS Nexus*. 2(9):pgad286. doi:10.1093/pnasnexus/pgad286
60 Mosleh M, Yang Q, Zaman T, Pennycook G, Rand DG. 2022. Analysis: trade-offs between reducing misinformation and politically-balanced enforcement on social media.130. *PsyArXiv*, Preprint, doi:10.31234/osf.io/ay9q5
61 Fan H, *et al.* 2021. Social media toxicity classification using deep learning: real-world application UK brexit. *Electronics*. 10:1332.
62 Awal MR, Cao R, Mitrovic S, Lee RK-W. 2020. On analyzing antisocial behaviors amid COVID-19 pandemic. *Arxiv Preprint*. 1–15. https://arxiv.org/abs/2007.10712
63 Petrizzo Z. 2021. Jason Miller's 'free speech' social media platform Gettr boots white nationalist [accessed 2023 Mar 16]. https://www.thedailybeast.com/jason-millers-free-speech-social-media-platform-gettr-boots-white-nationalist-nicholas-fuentes
64 Ng LHX, Cruickshank I, Carley KM. 2021. Coordinating narratives and the capitol riots on Parler. *arXiv, arXiv*:2109.00945, preprint: not peer reviewed.
65 Munn L. 2021. More than a mob: Parler as preparatory media for the U.S. Capitol stormingAuthors. *First Monday*. 26(3):1–16. doi:10.5210/fm.v26i3.11574
66 McGraw M. 2023. Bannon on Brazil riots: 'I'm not backing off 1 inch [accessed 2023 Jul 24]. https://www.politico.com/news/2023/01/09/bannon-brazil-riots-trump-00077155
67 Wendling M. 2023. How Trump's allies stoked Brazil congress attack [accessed 2023 Jun 26]. https://www.bbc.co.uk/news/world-us-canada-64206484
68 Tucker JA, Theocharis Y, Roberts ME, Barberá P.2017. From liberation to turmoil: social media and democracy. *J Democr*. 28:46–59.
69 Persily N, Tucker JA. 2020. *Social Media and Democracy*. New York: Cambridge University Press.
70 Observatory SI. 2021. Gogettr [accessed 2023 Feb 6]. https://github.com/stanfordio/gogettr
71 Morse J. 2021. Gettr, that site for twitter rejects, is mad twitter won't let it import tweets [accessed 2023 Jan 27]. https://www.nbcnews.com/think/opinion/twitter-lacked-something-important-political-discourse-joe-rogan-found-it-ncna1287285
72 Guimarães SS, Reis JCS, Ribeiro FN, Benevenuto F. 2020. *Characterizing toxicity on Facebook comments in Brazil*. In: Proceedings of the Brazilian Symposium on Multimedia and the Web, WebMedia '20. New York (NY): Association for Computing Machinery. p. 253–260.
73 Sipka A, Hannak A, Urman A. 2022. Comparing the language of QAnon-related content on Parler, Gab, and Twitter. In: 14th ACM Web Science Conference 2022, WebSci '22. New York (NY): Association for Computing Machinery. p. 411–421.
74 Kumar D, Hancock J, Thomas K, Durumeric Z. 2022. *Understanding longitudinal behaviors of toxic accounts on reddit. Arxiv Preprint*, 1–27. https://arxiv.org/abs/2209.02533
75 Jigsaw. 2023. Google perspective API: attributes and languages [accessed 2023 Mar 16]. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=enUS