SocArXiv

Preprint : June 1, 2023 *RedURL/DOI GOES HERE*

Suspensions of prominent accounts minimally impact platform engagement

Kayla Duskin
Information School
University of Washington

Jevin West
Information School
University of Washington

Joseph Bak-Coleman
Craig Newmark Center
Columbia University

Abstract

Health-related misinformation online poses threats to individual well-being and undermines public health efforts. In response, many social media platforms have temporarily or permanently suspended accounts that spread misinformation, at the risk of losing traffic vital to platform revenue. Here we examine the impact on platform engagement following removal of six prominent accounts during the COVID-19 pandemic. Focused on those who engaged with the removed accounts, we find that suspension did not meaningfully reduce activity on the platform. Moreover, we find that removal of the prominent accounts minimally impacted the diversity of information sources consumed.

Keywords: misinformation, social media, platform moderation, disinformation, discontinuity analysis.

Misinformation online has become a major area of concern, given its potential for harm to democracy, climate change, the safety of minoritized groups, and public health of the Surgeon General et al. (2021). During the coronavirus pandemic, health-related misinformation ranged widely from dubious cures to false information about the safety and efficacy of vaccines and treatments. There is evidence that health misinformation can cause harm, as exposure to vaccine-related misinformation can increase vaccine hesitancy which puts individuals at risk of severe disease and death Pierri et al. (2022). Social media platforms have responded with a number of interventions, including labeling or removing false content, "virality circuit breakers", and providing users with easy access to accurate information. Most major platforms likewise enacted community guidelines, violations of which can result in various forms of sanctioning. In extreme cases, accounts that repeatedly spread health-related misinformation were banned from using the platform entirely [1].

[1] https://web.archive.org/web/20210720070716/https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy

However, platforms may be reluctant to take action on highly followed accounts if doing so causes their audience to disengage with the service. To the extent that this is true—or perceived to be—there will be a poor alignment between platform incentives and public health goals. Here we address these questions by examining the activity of the users most likely to be impacted by account removal—those that were highly engaged with the sanctioned account prior to its suspension.

Previous work has shed light on deplatforming in the context of hate speech and conspiracy theories. A case study on the suspension of Twitter users known for their offensive speech showed that suspension decreased the toxicity and volume of tweets by the user's followers Shagun Jhaver, Christian Boylston, Diyi Yang, Amy Bruckman (2021). Other research on the migration of suspended users from mainstream platforms to less moderated platforms found that while the users' speech became more toxic on these alternative platforms, the size of their audience decreased substantially Ali et al. (2021); Ribeiro et al. (2020). Additionally, it has been shown that banning a topic forum (e.g. a subreddit) decreases the activity and interactions of that forum's users in other communities on the platform Engel et al. (2022); Saleem and Ruths (2018) but that users can regroup on other platforms Monti et al. (2023).

In this work, we study to what extent the objectives of reducing misinformation and maintaining an engaged user base are at odds with one another. We find that in most cases the engaged followers of a user that has been suspended for spreading COVID misinformation do not significantly decrease their activity on topics unrelated to COVID. For COVID related tweets, we see more fluctuations in engagement. We additionally hypothesized that removal of an influential account would cause followers to more equally distribute their attention to other accounts. We did not find cohesive evidence for this, suggesting that the diversity of sources that followers engage with is not generally impacted by suspending influential accounts.

Results

We identified six prominent Twitter accounts that were suspended for violating Twitter's COVID misinformation policy. For each of the suspended users, we selected two groups of 50 twitter accounts: a highly engaged (*HE*) group and a moderately engaged (*ME*) group.

For each user in each of the six *HE* and *ME* groups, we fit a Bayesian Gaussian Process Negative Binomial Regression Discontinuity model, as in Fig 1 Ornstein and Duck-Mayr (2022). We consider the effect of each suspension on a single engaged follower to be the percent change in daily posts (tweets, retweets, quote tweets, and replies) between the model's estimate of daily posts if no suspension had occurred and the observed posting frequency.

Engagement with Non-COVID Content

We find that the posting frequency in non-COVID topics did not meaningfully change for four out of the six suspensions for both the highly and moderately engaged followers, shown in Fig 2B, and Fig 2C respectively. For each prominent user, we found the change in posting frequency to be largely consistent between the *HE* and *ME* groups.

Engagement with COVID Related Content

Given that the suspensions we studied pertained to COVID misinformation, we repeated our

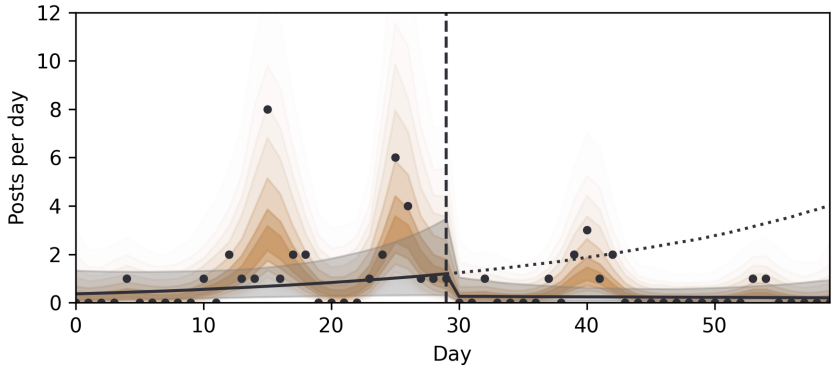

Figure 1: **Counterfactual account removal**. This figure provides an example of a Gaussian process regression discontinuity model for a single user. Points indicate the observed number of tweets (including retweets) on a given day, and the vertical line indicates the day of the corresponding suspended users' account removal. The shaded orange area indicates posterior predictive distribution of our model, with transparency indicating the 50 to 97% credible regions. The black line and shaded grey region indicate the estimated average rate of posting and 94% credible interval for the discontinuity model. The dotted line indicates the counterfactual predictions of the model.

analysis specifically on tweets on the topic of COVID or the COVID vaccine. Here we found that the suspension of the prominent user had more pronounced and heterogeneous effects. The *HE* and *ME* groups showed an increase in COVID-related tweeting for one out of the six suspensions, while four of the suspensions resulted in a meaningful decrease in posting for the *HE* groups and two showed negative change for the *ME* group as in Fig 1D and Fig 1E.

Information Diversity

To evaluate change in the diversity of content sources, we calculated the Shannon diversity Shannon (1948) of unique accounts retweeted by the engaged follower groups before and after the prominent users' suspensions. Again, for both *HE* and *ME* followers, we found little change in information diversity following suspension of prominent accounts. Combined, these findings suggest that suspension of a prominent account does little to impact the distribution and breadth of content interacted with by their followers.

Discussion

Under the scrutiny of the public and advertisers alike, social media platforms must weigh their decisions carefully when choosing their approach to content moderation. The recent Digital Services Act[2], recently passed in the European Union, will put additional pressure on these decisions.

This work finds that the removal of individual accounts does not consistently decrease the participation of that account's followers. From one viewpoint this could be seen as a discouraging sign for those wanting to use this as a strategy for mitigating harmful misinformation. Indeed, these results suggest that the community of vaccine-skeptical users is robust to the

[2]https://www.wilsoncenter.org/article/eus-digital-services-act-confronts-silicon-valley

de-platforming of individual influencers. However, from another angle this finding indicates that a platform choosing to suspend individual users may not lose their audience to the degree that might have previously been predicted. This suggests that de-platforming could serve as a tool to remove the most troubling content from online discourse without impacting the community as a whole.

Changes in posting frequency were inconsistent across individual suspended users. For one user, there was an increase in posting for both highly engaged and moderately engaged users while in the other suspended users change was negative or close to zero. What might be driving these differences? The volatility of larger trends regarding the topic of COVID could provide an explanation. In consulting CDC US COVID case counts, we observed that in the case of the suspended user whose engaged followers increased in posting volume, COVID case counts were on the rise while in other cases they were stagnant or decreasing. Further work is necessary to formally assess whether case counts played a larger role in online behavior than the removal of accounts.

When comparing our results to those of Shagun Jhaver, Christian Boylston, Diyi Yang, Amy Bruckman (2021), we saw no consistent decrease in followers' activity, while the prior study found a decrease in posting activity. This prior work defined the group of supporters via a more narrow definition of content specifically related to the deplatformed user. In contrast, this study examined content that dominated the headlines both in traditional media and social media as well as total platform engagement. We find a similar contrast with work that examined the effects of disbanding entire communities as in Saleem and Ruths (2018) and Engel et al. (2022).

This study looks at volume rather than a detailed content analysis. We tally the daily number of tweets, retweets, quotes and replies. We separate COVID and non-COVID tweets, but a more detailed, longitudinal analysis of content would need to look beyond this broad categorization. Our findings show little drop in posting activity following a de-platforming event. It is therefore possible that these suspensions mitigate harmful content from the suspended users without decreasing overall activity levels on the platform. The lack of meaningful change in shannon diversity of retweet sources also implies that deplatforming does not radically change the breadth of content consumed by their followers.

Moreover, our analysis is limited to those who engage with content via retweet, reply, or quote tweet. This does not take into account those that may read content but not engage with it directly. It is possible the effects on this group of users are wholly different in ways that are challenging to study given our access to platform data. Additionally, this analysis is specific to small and moderate influencers. The results could be different for the most prominent users, such as those with millions of followers as opposed to several hundred thousand. As with keystone species in biological systems Mills et al. (1993), there may be a tipping point when removing certain species or when removing a certain number of small to moderate users.

Overall, our study suggests that removal of prominent accounts sharing misinformation on a given topic may redirect followers' attention elsewhere, albeit with some changes in overall engagement. These results highlight potential trade-offs faced by platforms considering suspensions which may lead to perverse incentives to retain problematic users in order to maximize engagement. This study is not meant to be a policy recommendation for account removal. We do not encourage or discourage account removal, but we do hope the results can inform future policy. Future work is needed to evaluate whether these findings occur across

a broader range of topics and platforms, and to evaluate impacts on users that read but do not engage with content.

Data Collection

We identified six prominent Twitter users suspended from Twitter's platform for sharing COVID-19 related health-misinformation during 2021 (See SI). Using a dataset of tweets about COVID19 collected in real time, we identified highly engaged and moderately engaged followers of these six accounts. The *HE* group was defined as the 50 accounts that most frequently retweeted tweets about COVID posted by one of our prominent, suspended accounts before their suspension. To define the *ME* groups, we ranked each account that retweeted the suspended account more than once in the 30 days preceding its suspension by the number of retweets. We randomly sampled accounts that fell between the median and third quartile, in order to represent users who were engaged, but less so than the *HE* group. Using data collected and stored in real time, we were able to identify accounts which had been engaging with the account prior to its suspension. We then used the Twitter API to augment this data to get all tweets by all members of both groups, culminating in 2.5 million total tweets.

Statistical Models

For each user in the *HE* or *ME* group, we fit a Bayesian Gaussian Process (GP) Regression Discontinuity Model (Fig: 1, SI Methods). Briefly, this model decomposes our time-series into a mean and a variance function. The mean function characterizes the average rate of posting for engaged follower accounts before prominent account removal and the change after removal (Fig 1A) Ornstein and Duck-Mayr (2022). The variance function characterizes the (perhaps non-linear) variation around the mean. This allows us to account for the bias originating from temporal patterns of posting that would violate distributions of normality associated with typical regression discontinuity models. Changes in Shannon diversity were estimated using Bayesian Beta and Gamma regressions with weakly-informative priors. Full models are provided in the code.

0.1. Data and Code Availability

Data in the form of tweet ids and user ids and code necessary to reproduce our results available on https://github.com/kduskin/TwitterSuspendedUsers.

0.2. Acknowledgements

K.D. would like to acknowledge support by the National Science Foundation Graduate Research Fellowship under Grant No DGE-2140004. J.D.W would like to acknowledge the Knight Foundation and the NSF SaTC Grant (2120496) for supporting this research. J.B. would like to acknowledge support from Craig Newmark and the Berkman Klein Center's Institute for Rebooting Social Media.

References

Ali, S., Saeed, M. H., Aldreabi, E., Blackburn, J., De Cristofaro, E., Zannettou, S., and Stringhini, G. (2021). Understanding the effect of deplatforming on social networks. In *13th ACM Web Science Conference 2021*, WebSci '21, pages 187–195, New York, NY, USA. Association for Computing Machinery.

Engel, K., Hua, Y., Zeng, T., and Naaman, M. (2022). Characterizing reddit participation of users who engage in the qanon conspiracy theories. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22.

Mills, L. S., Soulé, M. E., and Doak, D. F. (1993). The keystone-species concept in ecology and conservation. *BioScience*, 43(4):219–224.

Monti, C., Cinelli, M., Valensise, C., Quattrociocchi, W., and Starnini, M. (2023). Online conspiracy communities are more resilient to deplatforming.

of the Surgeon General, O. et al. (2021). Confronting health misinformation: The us surgeon general's advisory on building a healthy information environment [internet].

Ornstein, J. and Duck-Mayr, J. (2022). Gaussian process regression discontinuity. *Working Paper*.

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K.-C., Flammini, A., Menczer, F., and Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Sci. Rep.*, 12(1):5966.

Ribeiro, M. H., Jhaver, S., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., and West, R. (2020). Do platform migrations compromise content moderation? evidence from r/The_Donald and r/incels.

Saleem, H. M. and Ruths, D. (2018). The aftermath of disbanding an online hateful community.

Shagun Jhaver, Christian Boylston, Diyi Yang, Amy Bruckman (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, pages 379–423.

Affiliation:

Kayla Duskin
University of Washington
Seattle, Washington, USA
E-mail: kduskin@uw.edu

SocArXiv Website https://socopen.org/
SocArXiv Preprints https://osf.io/preprints/socarxiv

Preprint *Submitted:* June 1, 2023
URL/DOI GOES HERE *Accepted:* June 1, 2023

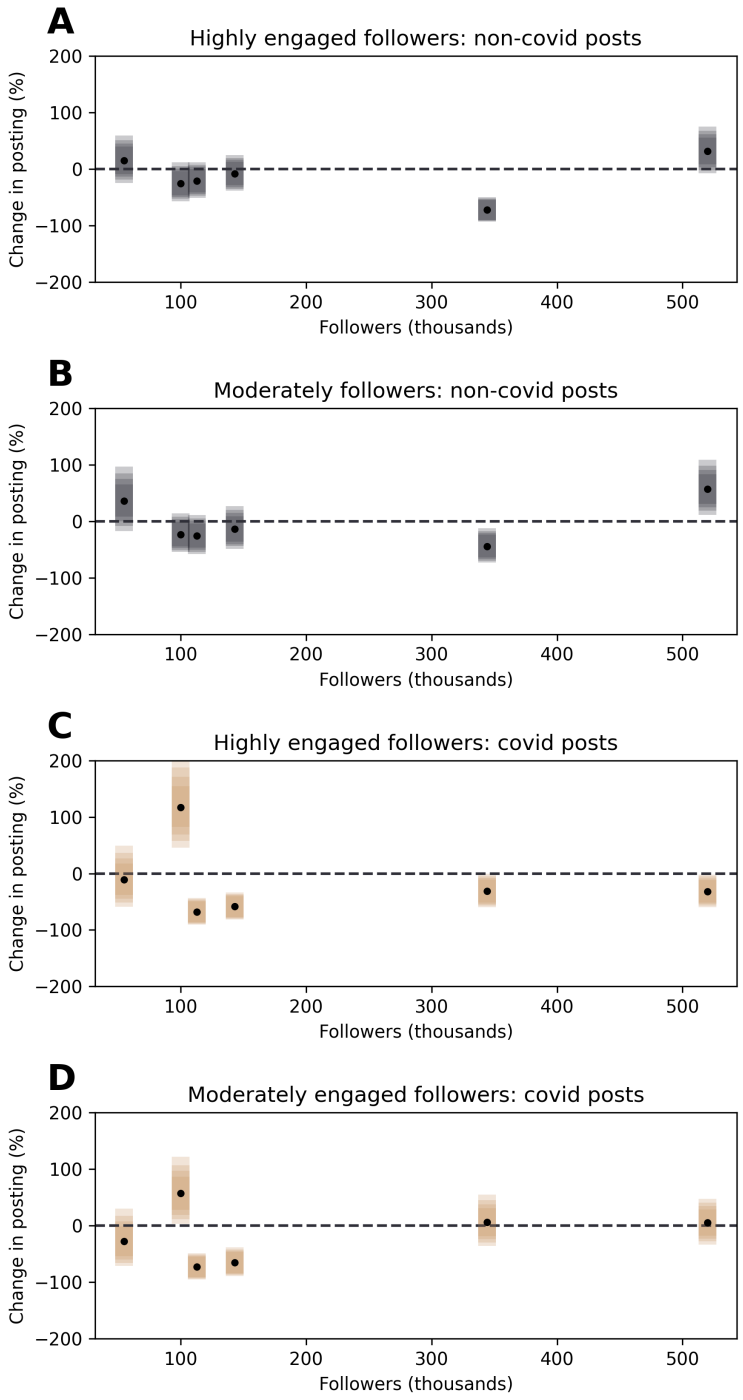

Figure 2: **Activity of highly and moderately engaged followers for non-COVID and COVID related tweets.** A) Posterior estimates of the relative change in post frequency for highly engaged followers one month after the suspension. Individual bars correspond to the activity of engaged followers for a given suspended account B) As in A but for median-engagement followers. C) For highly-engaged followers, indicating their change in COVID-related posting one month after the suspension event. D) As in C but for median-engaged followers. Higher values indicate more diverse engagement. Shaded regions in all panels indicate the 50, 75, 98, 94, and 97% credible intervals.

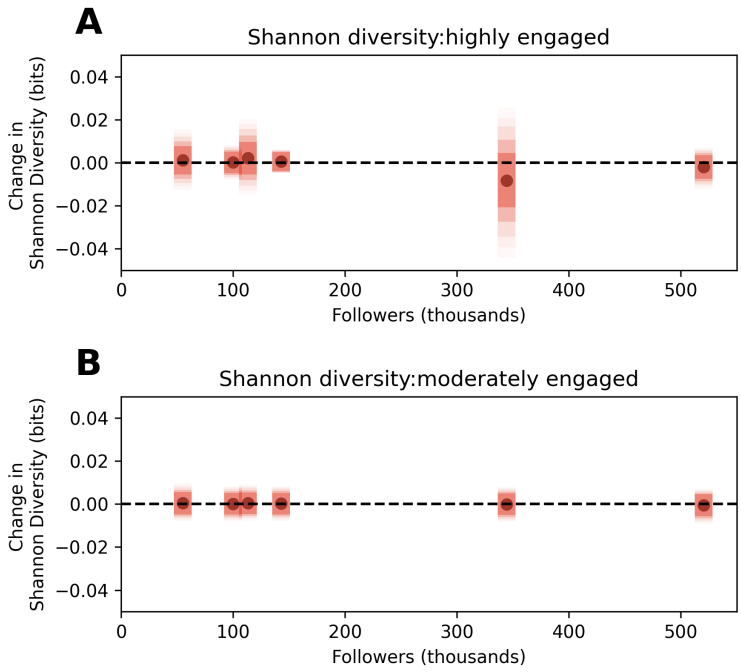

Figure 3: Information Diversity Effects. A) Change in Shannon Diversity Index in the month following suspension for highly engaged followers; higher values indicate more diverse engagement. B) As in A but for median engagement followers. Shaded regions indicate the 50, 75, 89, 94, and 97% credible regions.