

Reddit Community Interventions and Cross-Platform Response: Migration and Policy-Response Discourse on Voat

Shahan Ahmed

Montclair State University
ahmeds48@montclair.edu

Abstract. Platform moderation can disrupt online communities, but the responses that follow do not necessarily take the form of direct migration. Communities may relocate, reorganize, discuss moderation decisions, express censorship concerns, or continue activity on alternative platforms without clear evidence of user-level movement.

This study examines Reddit community bans and quarantines and their relationship with activity and migration-related discourse on Voat. Using data from the Multi-Platform Aggregated Dataset of Online Communities (MADOC), we analyze interventions affecting FatPeopleHate, GreatAwakening, MillionDollarExtreme, and CringeAnarchy. We combine computational discourse screening with a stratified independent human-validation procedure and population-adjusted estimates across 90-day pre- and post-intervention windows.

The results reveal heterogeneous cross-platform responses. FatPeopleHate experienced a substantial increase in overall Voat activity while the relative prevalence of policy-response and broader relevant discourse declined. CringeAnarchy showed declines in direct migration and policy-response discourse following its permanent ban, although some estimates were sensitive to uncertain human labels. GreatAwakening instead exhibited a modest increase in policy-response discourse without clear evidence of increased direct migration.

These findings suggest that platform interventions do not produce a uniform migration process. Overall destination-platform activity and migration-related discourse may move in different directions, highlighting the importance of distinguishing behavioral migration from broader discussion of moderation, censorship, governance, and community rebuilding.

Keywords: Deplatforming · Content moderation · Platform governance · Cross-platform migration · Reddit · Voat

1 Introduction

Online platforms increasingly use community-level moderation to address harassment, hate speech, misinformation, and other harmful behavior. Interventions range from visibility restrictions and quarantines to permanent community bans. Prior research shows that these actions can reduce harmful activity on the source

platform [3,2]. Moderation, however, may also prompt users and communities to relocate, fragment, rebuild elsewhere, or discuss censorship and platform governance on alternative platforms [5,7,1].

This creates an important measurement problem: *discussion about migration is not equivalent to migration itself*. References to Reddit, Voat, bans, censorship, rebuilding, or relocation may indicate a collective response to moderation without demonstrating that identifiable users actually moved between platforms. Likewise, destination-platform activity can increase sharply even while migration- or governance-related discourse becomes a smaller share of the conversation. Distinguishing overall activity from different forms of intervention-related discourse is therefore necessary for interpreting cross-platform responses.

We examine Reddit interventions involving FatPeopleHate, GreatAwakening, MillionDollarExtreme, and CringeAnarchy and their corresponding activity and discourse on Voat. Using the Multi-Platform Aggregated Dataset of Online Communities (MADOC) [6], we analyze permanent bans as well as the quarantine and later permanent ban of CringeAnarchy.

Our central research question is:

How do community bans and quarantines correspond with direct migration discourse, policy-response discourse, and broader intervention-related activity on an alternative platform?

We combine event-centered activity analysis with computational text screening and independent human validation. For comparable cases, we examine 90-day pre- and post-intervention windows. Automated indicators screen for references to migration, Reddit and Voat, intervention, censorship or free speech, community rebuilding, and governance backlash. Because narrow migration patterns can misclassify general platform discussion, we validate the measures using a stratified manually coded sample and population-adjusted weights.

The study contributes in three ways. First, it separates explicit direct-migration discourse from broader policy and platform responses. Second, it compares multiple interventions using a common event-window framework while preserving cross-community heterogeneity. Third, it combines scalable computational screening with weighted human validation rather than treating keyword matches as definitive evidence of cross-platform movement.

The results reveal no uniform post-intervention trajectory. FatPeopleHate shows a sharp increase in overall Voat activity while the prevalence of policy-response and broader relevant discourse declines. CringeAnarchy shows post-ban declines in direct-migration and policy-response discourse, although some estimates are sensitive to uncertain labels. GreatAwakening instead shows a modest increase in policy-response discourse without clear evidence of increased direct migration. Overall activity and migration-related discourse therefore need not move in the same direction.

2 Related Work

Community-level moderation can reduce problematic behavior on the platform imposing the intervention, although effects vary across communities [9,4]. Chandrasekharan et al. found that Reddit’s 2015 bans of FatPeopleHate and Coon-Town were followed by user departures and large reductions in hate-speech use among users who remained [3]. Across 15 banned subreddits, Trujillo et al. similarly found heterogeneous changes in activity and community-specific language [10]. Interventions short of removal can produce different effects: Reddit quarantines substantially reduced recruitment into targeted communities while showing limited changes in the behavior of existing members [2].

A related literature examines responses beyond the source platform. Horta Ribeiro et al. studied The_Donald and Incels after migration from Reddit to dedicated websites and found declines in participation alongside evidence that some content characteristics could become more extreme [5]. Monti et al. directly examined FatPeopleHate and GreatAwakening migration from Reddit to Voat, finding that both partially migrated but that GreatAwakening was more resilient and more likely to reconstruct prior social connections [7]. Migration can also continue across successive platforms, as illustrated by QAnon-related movement from Reddit to Voat and later from Voat to Poal [8]. Cross-platform responses following the January 6 deplatforming likewise differed across alternative platforms in attention, engagement, and discourse [1].

Together, these studies show that moderation can correspond with departure, relocation, fragmentation, or reorganization rather than a single uniform response. Our study complements user-matching approaches by distinguishing what can be inferred from discourse itself. We separate explicit direct-migration discourse from broader policy-response discourse—including intervention, censorship, governance, and rebuilding discussion—and from general relevant platform discourse. This prevents references to an alternative platform from being treated automatically as evidence that identifiable users migrated.

3 Data and Methods

3.1 Data, Event Windows, and Discourse Measures

We use the Multi-Platform Aggregated Dataset of Online Communities (MADOC), which provides harmonized records from multiple online platforms, including Reddit and Voat [6]. The analysis focuses on Voat communities associated with Reddit communities that experienced a major platform intervention.

Five intervention events are considered: the permanent ban of FatPeopleHate on June 10, 2015; the permanent ban of MillionDollarExtreme on September 10, 2018; the permanent ban of GreatAwakening on September 12, 2018; the quarantine of CringeAnarchy on September 27, 2018; and the permanent ban of CringeAnarchy on April 25, 2019. The CringeAnarchy quarantine and permanent ban are treated as distinct interventions.

Table 1. Intervention cases and analytical sample sizes. Counts refer to Voat interactions within the corresponding 90-day event windows.

Intervention	Date	Type	Pre	Post	Role
FatPeopleHate	2015-06-10	Ban	180	46,750	Primary
MillionDollarExtreme	2018-09-10	Ban	0	6,922	Post only
GreatAwakening	2018-09-12	Ban	3,727	36,051	Primary
CringeAnarchy quarantine	2018-09-27	Quarantine	5	30	Descriptive
CringeAnarchy ban	2019-04-25	Ban	369	2,048	Primary

Table 1 summarizes the analytical windows. FatPeopleHate, GreatAwakening, and the permanent CringeAnarchy ban contain sufficient observations for pre-post comparison. MillionDollarExtreme has no observations in the corresponding pre-intervention window and is therefore described only after the intervention. The CringeAnarchy quarantine has only five pre-intervention interactions and is treated as descriptive rather than as a primary comparative case.

For each intervention, we construct an event-centered window using the official intervention date as day zero. The pre-intervention period covers the 90 calendar days immediately preceding the intervention (days -90 through -1), while the post-intervention period covers the intervention date and the following 89 days (days 0 through $+89$). This produces comparable 90-day windows while preventing the intervention date itself from being included in the pre-period.

The analysis is observational. Pre-post differences are interpreted as temporal changes associated with an intervention rather than causal effects. Other contemporaneous events may also influence activity and discourse within the analyzed communities.

We use computational text screening to identify several forms of intervention-related discourse. The screening procedure searches the text of Voat interactions for references to Reddit, Voat, platform interventions, censorship and free speech, migration or relocation, community rebuilding, and governance backlash.

These indicators are combined into three nested analytical outcomes. *Direct migration* represents the narrowest category and requires language that explicitly or strongly implies movement between platforms, such as statements about moving from Reddit to Voat, Reddit refugees, or migration following a ban. *Policy-response discourse* is broader and includes direct migration as well as discussion connecting Reddit or Voat with moderation interventions, censorship, free speech, rebuilding, or related governance responses. Finally, *relevant discourse* is the broadest measure and includes any substantive reference captured by the intervention-related screening system.

The three outcomes are intentionally distinguished because a reference to Reddit, a ban, or censorship does not by itself demonstrate cross-platform migration. Automated screening is therefore used primarily to identify candidate discourse rather than to establish behavioral movement of identifiable users.

3.2 Manual Validation and Population Weighting

Because intervention-related discourse is relatively rare and automated keyword patterns can generate both false positives and false negatives, we use stratified human validation. Candidate interactions were divided within each event-period combination into four mutually exclusive sampling strata: (1) strict migration matches, (2) broader policy-response matches that were not strict migration matches, (3) other relevant-discourse matches, and (4) interactions with no detected signal.

Up to 25 interactions were sampled from each available stratum using a fixed random seed, producing 609 manually coded records. A single independent human coder evaluated each interaction for overall relevance, direct migration, and a primary discourse category. The primary categories were direct cross-platform migration, intervention discussion, censorship or free speech, community rebuilding, governance backlash, platform reference only, unrelated, and ambiguous. Direct migration required an explicit or clearly implied movement between platforms. Uncertain judgments were recorded as U rather than being forced into a binary category.

Because the validation design deliberately oversampled interactions containing automated signals, raw proportions in the validation sample do not represent population prevalence. Each manually coded observation was therefore assigned an inverse sampling-fraction weight

$$w_h = \frac{N_h}{n_h},$$

where N_h is the number of population interactions in an event-period sampling stratum and n_h is the number manually sampled from that stratum. Weighted estimates reconstruct event-period population rates for direct migration, policy-response discourse, and relevant discourse.

For the three intervention cases with adequate pre- and post-period data, we report the difference in weighted prevalence as

$$\Delta = p_{\text{post}} - p_{\text{pre}},$$

expressed in percentage points. Ninety-five percent confidence intervals account for the stratified validation design and finite population corrections. These intervals quantify uncertainty arising from the manual-validation sampling design; they should not be interpreted as causal confidence intervals.

The primary analysis excludes uncertain human judgments from binary classification. As a robustness check, we also calculate extreme bounds in which all uncertain observations are alternatively treated as negative or positive. This sensitivity analysis is used to identify findings that depend heavily on the small number of ambiguous observations, particularly when those observations carry large sampling weights.

4 Results

4.1 Validation of the Discourse Measures

Manual validation showed substantial differences across the three computational discourse measures. After accounting for the stratified sampling design, the broad relevant-discourse detector achieved high recall (0.974) with an F1 score of 0.704. The policy-response detector was substantially more conservative, with precision of 0.593, recall of 0.378, and an F1 score of 0.462.

The strict direct-migration detector performed poorly as a standalone classifier. Its weighted precision was 0.201, recall was 0.130, and F1 was 0.158. Inspection of disagreements showed that strict patterns missed contextually expressed migration while also occasionally classifying intervention or rebuilding language as direct movement. Consequently, automated direct-migration matches are not treated as definitive evidence of migration. The substantive pre-post estimates below instead rely on the weighted human-coded classifications.

These results indicate that computational screening is useful for retrieving broad intervention-related discourse, but increasingly narrow semantic claims require human validation.

4.2 Cross-Event Pre-Post Changes

Figure 1 shows overall Voat activity trajectories around the three comparable permanent bans, while Figure 2 summarizes the weighted human-validated percentage-point changes and confidence intervals. Together, these measures distinguish changes in overall activity volume from changes in the composition of migration- and policy-related discourse.

FatPeopleHate displayed the clearest reduction in intervention-related discourse. Direct-migration discourse decreased from 3.81% before the ban to 0.56% afterward, a change of -3.25 percentage points (pp; 95% CI $[-4.04, -2.46]$). The decrease was considerably larger for policy-response discourse, from 10.99% to 1.50% ($\Delta = -9.50$ pp, 95% CI $[-10.79, -8.21]$), and for broader relevant discourse, from 17.80% to 2.62% ($\Delta = -15.18$ pp, 95% CI $[-17.29, -13.07]$). The decreases in policy-response and relevant discourse remained directionally robust under the extreme treatment of uncertain human labels. The direct-migration estimate was more sensitive to those uncertain observations and is therefore interpreted more cautiously.

CringeAnarchy showed a broadly similar directional pattern after its permanent ban. Direct-migration discourse declined from 6.33% to 0.81% ($\Delta = -5.52$ pp, 95% CI $[-8.36, -2.67]$), while policy-response discourse declined from 13.97% to 4.25% ($\Delta = -9.72$ pp, 95% CI $[-14.08, -5.36]$). Broader relevant discourse also declined, but its confidence interval crossed zero ($\Delta = -5.76$ pp, 95% CI $[-15.67, 4.15]$). Moreover, the CringeAnarchy estimates were sensitive to alternative assignments of uncertain human labels. They therefore provide evidence of a post-intervention decline in direct-migration and policy-response discourse, but with weaker robustness than the principal FatPeopleHate results.

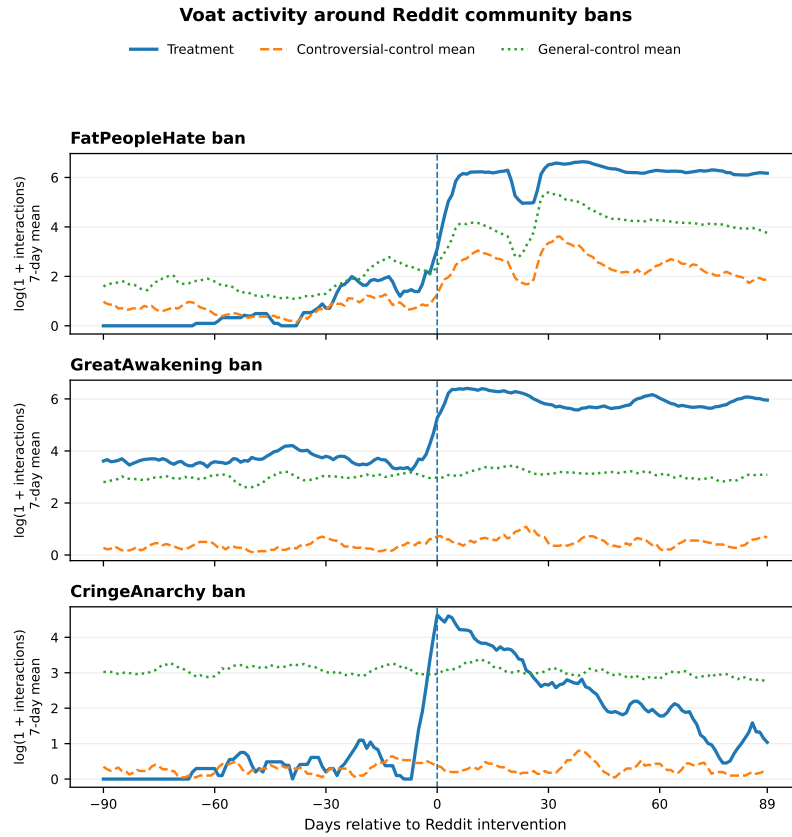


Fig. 1. Voat activity around three Reddit community bans. Lines show 7-day moving averages of $\log(1 + \text{daily interactions})$; the vertical dashed line marks the intervention date.

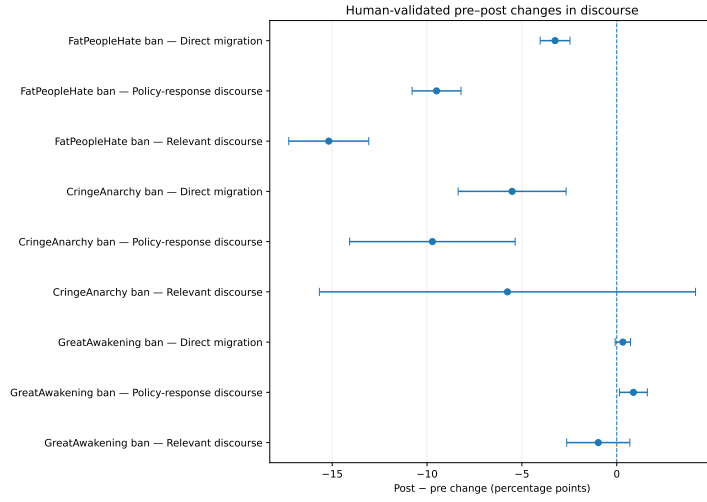


Fig. 2. Human-validated pre-post discourse changes. Points show weighted percentage-point changes with 95% design-based confidence intervals.

GreatAwakening followed a different trajectory. Policy-response discourse increased from 0.43% to 1.30%, corresponding to a 0.88 pp increase (95% CI [0.14, 1.61]). Direct-migration discourse increased only from 0.11% to 0.43%, with a confidence interval spanning zero ($\Delta = 0.32$ pp, 95% CI [-0.08, 0.73]). Relevant discourse decreased from 3.57% to 2.59%, but this difference was also uncertain ($\Delta = -0.97$ pp, 95% CI [-2.64, 0.69]). The positive policy-response estimate was sensitive to extreme treatment of uncertain labels and should therefore be interpreted as suggestive rather than definitive evidence of an increase.

4.3 Small and Incomplete Intervention Cases

The CringeAnarchy quarantine contained only five interactions in the 90-day pre-intervention window and 30 interactions in the post-window. Although weighted estimates can be calculated for these records, the extremely small population makes conventional comparative interpretation inappropriate. We therefore retain the quarantine only as descriptive evidence.

MillionDollarExtreme presents the opposite limitation: 6,922 interactions are available after the September 2018 ban, but no observations are available in the corresponding 90-day pre-period. Within the post-period, the weighted human estimates were 1.57% for direct migration, 3.26% for policy-response discourse, and 5.69% for relevant discourse. These values characterize post-intervention discourse but cannot establish a pre-post change.

5 Discussion

The cross-case results show that Reddit interventions were not followed by a uniform pattern of activity or migration-related discourse on Voat. Instead, the cases exhibit different combinations of destination activity, direct-migration discourse, and broader policy response. This heterogeneity is consistent with prior work showing that community responses to moderation depend on the characteristics of the affected community and its members [10,7].

FatPeopleHate provides the clearest illustration of why activity volume and discourse composition should be distinguished. Overall Voat activity increased sharply around the Reddit ban (Fig. 1), yet the weighted prevalence of policy-response and broader relevant discourse declined substantially. The result therefore does not indicate an overall contraction of the destination community. Rather, intervention- and migration-related discussion became a smaller share of a rapidly expanding stream of activity. A post-ban activity surge consequently should not be treated as equivalent to a sustained increase in migration discourse.

CringeAnarchy followed a different temporal pattern. Activity began increasing immediately before the permanent ban, peaked near the intervention, and subsequently declined. This pre-ban mobilization reinforces the need for caution in attributing the observed trajectory solely to the intervention. Human-validated direct-migration and policy-response discourse nevertheless declined between the 90-day windows, although these estimates were less robust to uncertain labels than the strongest FatPeopleHate findings.

GreatAwakening provides a contrasting case. Treatment-community activity increased substantially around the ban and remained elevated, while policy-response discourse showed a modest increase. Direct migration, however, did not show a clearly distinguishable increase. Thus, heightened discussion of platform governance or intervention does not necessarily provide evidence of proportional increases in direct cross-platform movement.

More generally, these results support separating at least three dimensions of cross-platform response: overall destination activity, explicit migration discourse, and broader policy or governance discussion. Automated screening was effective for retrieving broad candidate discourse but performed poorly for the narrow direct-migration category. Human validation is therefore particularly important when textual evidence is used to make claims about migration rather than general platform response.

6 Limitations and Conclusion

Several limitations qualify the findings. The analysis identifies migration-related discourse rather than matched cross-platform identities, so it cannot establish individual-level migration, and Voat is only one possible destination. The observational pre-post design may reflect concurrent events or existing trends, particularly for CringeAnarchy. Coverage is uneven: the quarantine is sparse, MillionDollarExtreme lacks a comparable pre-period, and a small number of uncertain records carry large sampling weights. Even so, the cross-case evidence

shows that destination activity, explicit migration discourse, and broader governance response can diverge following platform intervention; these dimensions should therefore be analyzed separately.

References

1. Buntain, C., Innes, M., Mitts, T., Shapiro, J.N.: Cross-platform reactions to the post-january 6 deplatforming. *Journal of Quantitative Description: Digital Media* **3** (2023). <https://doi.org/10.51685/jqd.2023.004>
2. Chandrasekharan, E., Jhaver, S., Bruckman, A., Gilbert, E.: Quarantined! examining the effects of a community-wide moderation intervention on reddit. *ACM Transactions on Computer-Human Interaction* **29**(4) (2022). <https://doi.org/10.1145/3490499>
3. Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., Gilbert, E.: 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) (2017). <https://doi.org/10.1145/3134666>
4. Cima, L., Trujillo, A., Avvenuti, M., Cresci, S.: The great ban: Efficacy and unintended consequences of a massive deplatforming operation on reddit. In: Companion Publication of the 16th ACM Web Science Conference. pp. 85–93. Association for Computing Machinery (2024). <https://doi.org/10.1145/3630744.3663608>
5. Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., West, R.: Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction* **5**(CSCW2) (2021). <https://doi.org/10.1145/3476057>
6. Mitrović Dankulov, M., Tomašević, A., Maletić, S., Anđelković, M., Vranić, A., Cvetković, D., Stupovski, B., Vudragović, D., Major, S., Bogojević, A.: Multi-platform aggregated dataset of online communities (MADOC). *Proceedings of the International AAAI Conference on Web and Social Media* **19**(1), 2529–2538 (2025). <https://doi.org/10.1609/icwsm.v19i1.35954>
7. Monti, C., Cinelli, M., Valensise, C., Quattrociocchi, W., Starnini, M.: Online conspiracy communities are more resilient to deplatforming. *PNAS Nexus* **2**(10), pgad324 (2023). <https://doi.org/10.1093/pnasnexus/pgad324>
8. Papasavva, A., Mariconti, E.: Waiting for q: An exploration of qanon users' online migration to poal in the wake of voat's demise. In: *Proceedings of the 16th ACM Web Science Conference*. pp. 250–261. Association for Computing Machinery (2024). <https://doi.org/10.1145/3614419.3644021>
9. Thomas, P.B., Riehm, D., Glenski, M.F., Weninger, T.: Behavior change in response to subreddit bans and external events. *IEEE Transactions on Computational Social Systems* **8**(4), 809–818 (2021). <https://doi.org/10.1109/TCSS.2021.3061957>
10. Trujillo, M., Rosenblatt, S., de Anda Jáuregui, G., Moog, E., Samson, B.P.V., Hébert-Dufresne, L., Roth, A.M.: When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban. In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. pp. 164–178. Association for Computational Linguistics (2021). <https://doi.org/10.18653/v1/2021.woah-1.18>