Understanding Online Migration Decisions Following the Banning of Radical Communities

Giuseppe Russo
Chair of System Design, ETH Zurich
Switzerland
russog@ethz.ch

Manoel Horta Ribeiro
Data Science Lab, EPFL
Switzerland
manoel.hortaribeiro@epfl.ch

Giona Casiraghi
Chair of System Design, ETH Zurich
Switzerland
gcasiraghi@ethz.ch

Luca Verginer
Chair of System Design, ETH Zurich
Switzerland
lverginer@ethz.ch

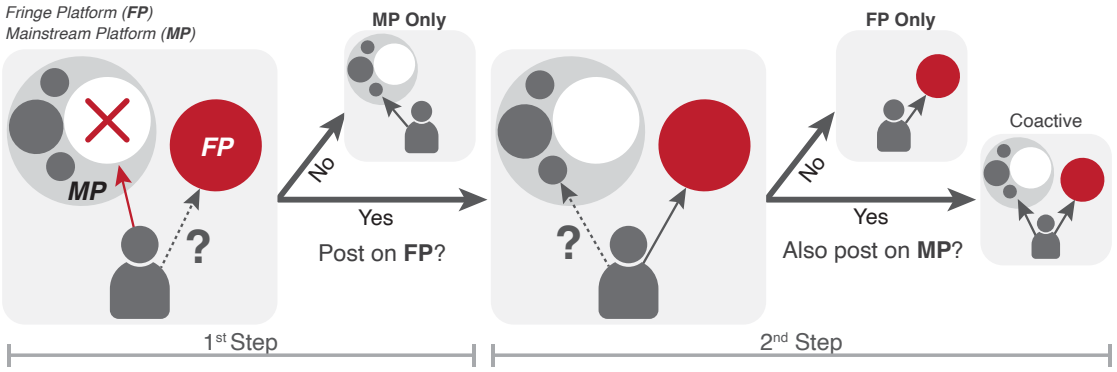

Figure 1: When a community is banned from a mainstream platform for harmful conduct (indicated by an × in the figure), users have to decide whether to (i) migrate to fringe platforms where the community migrated toward; and (ii) remain active in other communities in the mainstream platform. In this paper, we study factors associated with these two decisions.

ABSTRACT

The proliferation of radical online communities and their violent offshoots has sparked great societal concern. However, the current practice of banning such communities from mainstream platforms has unintended consequences: (i) the further radicalization of their members in fringe platforms where they migrate; and (ii) the spillover of harmful content from fringe back onto mainstream platforms. Here, in a large observational study on two banned subreddits, `r/The_Donald` and `r/fatpeoplehate`, we examine how factors associated with the RECRO radicalization framework relate to users' migration decisions. Specifically, we quantify how these factors affect users' decisions to post on fringe platforms and, for those who do, whether they continue posting on the mainstream platform. Our results show that individual-level factors, those relating to the behavior of users, are associated with the decision to post on the fringe platform. Whereas social-level factors, users' connection with the radical community, only affect the propensity to be coactive on both platforms. Overall, our findings pave the way for evidence-based moderation policies, as the decisions to migrate and remain coactive amplify unintended consequences of community bans.

WebSci '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0089-7/23/04…$15.00
https://doi.org/10.1145/3578503.3583608

ACM Reference Format:
Giuseppe Russo, Manoel Horta Ribeiro, Giona Casiraghi, and Luca Verginer. 2023. Understanding Online Migration Decisions Following the Banning of Radical Communities. In *15th ACM Web Science Conference 2023 (WebSci '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3578503.3583608

1 INTRODUCTION

Online platforms enforce strict moderation policies to prevent the spread of content deemed harmful [12, 31]. Consequently, they ban communities breaching their guidelines [4], often resulting in the migration of these communities to fringe, alternative platforms with little to no content moderation [10].

Previous work suggests that the migration of banned communities radicalizes their members [1, 47]. For example, after the prominent Reddit community `r/The_Donald` migrated to a self-hosted website, *thedonald.win*, its users became significantly more toxic [16]. Further, toxic behavior accepted on fringe platforms spills back onto the mainstream platforms through coactive users, i.e., users that remain active on both [36]. For example, users migrating to *thedonald.win and* continuing to post on Reddit became more toxic also on the latter. Therefore, users of banned communities decide (i) whether to participate on the fringe platform (*migration decision*) and (ii) whether to participate in both platforms or abandon the original one (*coactivity decision*). Understanding the factors driving these decisions informs stakeholders of the externalities of community bans, paving the way for more evidence-based moderation policies.

To identify these factors, we build on RECRO, a theoretical model for internet-mediated radicalization proposed by Neo [26]. RECRO defines five phases of radicalization: Reflection, Exploration,

Connection, Resolution, and Operational. Only the first three phases describe online activity. The latter two describe real-world actions that are not observed in online data. For this reason, we focus on studying factors related to Reflection, Exploration, and Connection. *Reflection factors* describe needs and vulnerabilities that render individuals more receptive to alternate belief systems. *Exploration factors* quantify how individuals make sense of information put forth by the radical community. Finally, *Connection factors* describe the influence of community members on each other. We hypothesize that factors associated with how users reflect, explore, and connect (REC) with radical communities explain their migration decisions following bans.

Our analysis confirms that REC factors are indicative of migration and coactivity decisions. Interestingly, the factors associated with each decision differ. While reflection factors correlate mostly with the decision to migrate to the fringe platform, Connection factors are mainly associated with the coactivity decision. Since Reflection describes an individual's online behavior while Connection describes an individual's social environment, we conclude that individual motives drive the decision to engage with the new platform, whereas social factors drive coactivity.

2 RELATED WORK

Antisocial online communities. Various online communities consistently engage in antisocial behavior [23], often harassing minorities and sympathizing with extremist ideologies [33, 40]. These communities have disproportionate influence over memes and news shared on the web [44]. Further, they have been closely associated with medical misinformation [46], conspiracy theories [39], and extremist ideologies [24].

Among those communities on Reddit, we consider `r/The_Donald` and `r/fatpeoplehate`. The subreddit `r/The_Donald` was created in June 2015 to support the then-presidential candidate Donald Trump's bid for the U.S. Presidential election. This community has been closely linked with the rise of the "alt-right" movement hosting racist, sexist and islamophobic discussions [22], and spreading conspiracy theories [28]. Flores-Saviaga et al. [9] have studied how active participants in `r/The_Donald` mobilized the community to engage in "political trolling." The subreddit `r/fatpeoplehate` was created in June 2015 to promote collective actions of body shaming and violence against overweight people. In 2015, Reddit banned `r/fatpeoplehate` following a newly-introduced policy to ban subreddits targeting and harassing specific groups [3].

Online Migration. Both `r/The_Donald` and `r/fatpeoplehate` have been "de-platformed," i.e., banned from Reddit for breaching their guidelines. In response to the banning, users of these communities migrated to fringe platforms where they continued their discussions. For example, users of `r/The_Donald` migrated *en masse* to *thedonald.win*, a self-hosted website where users of `r/The_Donald` could continue discussing and behaving as before. Similarly, users of `r/fatpeoplehate` migrated to the Reddit-like platform, *voat.co*, where they re-established their community.

Previous work has studied the effects of deplatforming, finding that after a ban, users reduce their activity on mainstream platforms [17], but also that users often migrate to other fringe platforms, where they become more toxic [1]. These studies imply that user migration, as an outcome of community deplatforming, might isolate users and expose them to more extreme content. Additionally, Russo et al. [36] identified a 'radicalization spillover' from fringe to mainstream platforms caused by users that migrated to the fringe but remained active on the mainstream platform (referred to as *coactive* users). Therefore, previous research concludes that online migration of radical communities yields consequences at community and platform levels. At the community level, migration to fringe platforms can push users to further radicalize and participate in real-world violent events. At the platform level, users that have decided to migrate but remain active on the mainstream platform, *coactive* users, can undermine the efficacy of moderation policies such as banning.

Online Radicalization. Radicalization is defined as the adoption of extreme political, social, or religious ideals and aspirations that reject or undermine the status quo of society (e.g., acceptance of differences), which can lead to violence to achieve these goals [6]. Multiple radicalization models have been proposed [20, 21]. However, they either ignore internet-mediated radicalization or focus on psychological predispositions [26]. The RECRO model proposed by Neo is a theoretical model for internet-mediated radicalization. It has been used to study anti-vaccination discussions [41] and engagement with conspiracy theories [29]. The RECRO model consists of five phases: Reflection, Exploration, Connection, Resolution, and Operational (RECRO). As noted by Neo [26], these phases may overlap and occur multiple times in the radicalization process. Therefore, we consider these phases simultaneously to include such an overlap feature of the RECRO model.

Similarly to Phadke et al. [29], we focus on the first three phases—Reflection, Exploration, and Connection—based on how individuals engage with radical online content. The **R**eflection phase describes an individual's emotional state, making them susceptible to radical narratives. Neo [26] qualitatively find that heightened emotions like aggressiveness and anxiety are linked to *reflection*. The **E**xploration phase describes how individuals consume content, look for new information, and get exposed to radical narratives. Finally, **C**onnection describes how interactions with members of radical communities influence the individual.

Relation to prior work. Previous work has found that community-level bans can have negative externalities: users can migrate to other, more toxic communities [16] and may cause 'radicalization spillovers' from fringe to mainstream platforms [36]. As these consequences arise mainly from user decisions to migrate and to remain coactive, we build upon the RECRO radicalization framework [26] to investigate the factors associated with these decisions.

3 MATERIALS AND METHODS

3.1 Data

To study migration decisions, we use data from two subreddits (`r/The_Donald` and `r/fatpeoplehate`; see Section 2) and the fringe platforms their users migrated *en masse* after they were banned (*thedonald.win* and *voat.co*) We collect the entire posting history relevant to the two communities on Reddit and the fringe platforms.

Reddit. Using the Push API [2], we collect the posts made on the two subreddits, starting six months before they were banned. Specifically, for `r/fatpeoplehate`, we collect from February 1, 2015, to August 1, 2015; for `r/The_Donald`, from November 11, 2019, to

February 26, 2020. For each subreddit, we also collect all contributing users' entire Reddit posting history.

Fringe Platforms. We obtain *thedonald.win* data using custom Web crawlers and *voat.co* data from Mekacher and Papasavva [25]. For each platform, we collect posts made in the 36 weeks around the ban.

We discard users with low activity to ensure that our analysis is not biased by users with low engagement within the analyzed communities (similarly to [37]). Precisely, we discard users who made less than ten posts on `r/The_Donald` or `r/fatpeoplehate` *before* the ban. Further, among those users that post on reddit after the ban we discard those that contributed with less that ten posts on the whole Reddit *after* the ban. Finally, we assume users were active on the fringe platforms after the ban only if they made at least five posts. Overall, we collect 2.5 million posts from `r/The_Donald` and `r/fatpeoplehate`. These, combined with the data obtained from the other 4,786 subreddits, yield a total of 91.2 million posts by ∼ 140,000 users (91,244 for `r/The_Donald`, 49,765 for `r/fatpeoplehate`). Moreover, we collect over 2.5 million posts by 38,510 users from *thedonald.win*, and 1.3 million posts from 26,223 users from *voat.co*.

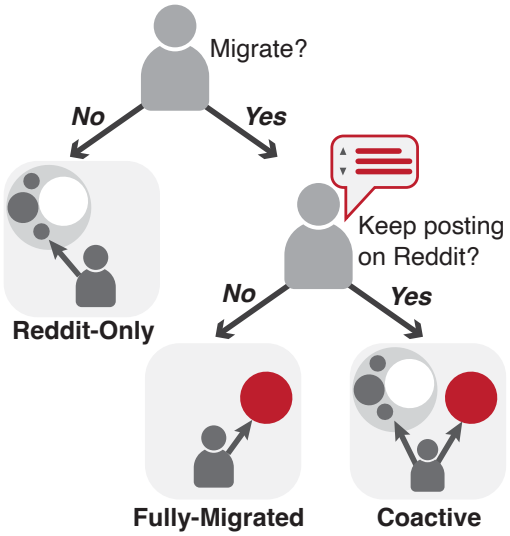

Figure 2: Mapping migration decisions to users' labels.

Users labeling. After the ban, users decide whether to post on the fringe platform or not. We label users that post on the fringe platform as *migrated* (MG). Those users that keep posting on Reddit but never post on the fringe platform are labeled as *Reddit-Only* (RO). In the second step of the migration, users that posted on the fringe platform may decide if posting on both platforms or to post exclusively on the fringe. We label those who continue posting on Reddit and the fringe platform as *coactive* (CA). Differently, users who stop posting on Reddit after the ban and post *exclusively* on the fringe platform are labeled as *Fully-Migrated* (FM). To track users across platforms, we apply exact string matching on their usernames, following previous works [16, 27]. This strategy emphasizes precision over recall: users who have changed their usernames during the migration are not considered migrated. We account for this fact while interpreting our results. Importantly, `r/The_Donald` had a system to facilitate username continuity across platforms [8]. In fig. 2, we show how these labels are mapped to the different migration decisions.

3.2 Modelling Migration

We investigate how radicalization factors drive migration decisions following community banning. To answer this question, we formalize platform migration as a two-step decision process (see fig. 1). In the first step, users become active on the fringe platform and start posting there. In the second step, users posting on the fringe platform choose whether to post on both platforms or to cease all activity on the mainstream platform. Previous work suggests that users explore fringe platforms and decide whether to participate based on a variety of reasons [27], e.g., the abundance of niche content in the mainstream platform and the permissive moderation style of fringe platforms.

We model the migration process with a Heckman two-stage regression [15]. The first stage models the propensity s_i of the user i to post on the fringe platform after the ban (first migration decision). The second stage models the likelihood that the user i remains coactive, accounting for their propensity s_i to post on the fringe platform (coactivity decision). For both stages, for a given user i, we consider various factors related to Reflection, Exploration, and Connection referred to as vectors $\mathbf{R_i}$, $\mathbf{E_i}$, and $\mathbf{C_i}$ (see section 4). Formally,

$$s_i \sim \gamma_0 + \overbrace{\gamma_1 \mathbf{R_i} + \gamma_2 \mathbf{E_i} + \gamma_3 \mathbf{C_i}}^{\text{REC variables}} + \gamma_4 \text{ER}_i \quad (1)$$

$$\text{CA}_i \sim \beta_0 + \underbrace{\beta_1 \mathbf{R_i} + \beta_2 \mathbf{E_i} + \beta_3 \mathbf{C_i}}_{\text{REC variables}} + \psi(s_i)\,, \quad (2)$$

where (2) is the main equation and (1) is the selection equation. Note that $\psi(s_i)$ accounts for the propensity s_i of user i to post on the fringe platform. Also, to improve the error estimates, we include an exclusion restriction, as suggested by Puhani [30]. The exclusion restriction is a variable affecting the selection propensity (s) but not the main outcome (CA_i). Precisely this is the variable ER in (1). In our case, the variable must affect the decision to migrate but only weakly affect coactivity. To this end, we choose *language coherence* (see section 4), measured as the ability to conform to the language used by the focal community. Continuing to use the language of the focal community may drive users to follow the group on the fringe platform. However, language usage does not preclude continued activity on the mainstream platform.

Importantly, if we were to model these stages without the Heckman correction, i.e., considering all users in the first stage and only migrating users in a second, separate regression, we would estimate the regression parameters incorrectly. Moreover, we would *not* answer the general question: *what is the probability of any user becoming coactive if their community was banned?* Instead, we would estimate which factors increase the probability that a migrated user (MG) becomes coactive (CA). The reason for this is selection bias. Indeed, we only observe coactivity for a subset of the population and not a random and representative sample of the banned community (see fig. 2).

4 OPERATIONALIZING RECRO

We operationalize the **R**eflection, **E**xploration, and **C**onnection stages of the RECRO framework using language, activity, and interaction features. These are calculated on the *pre-banning* activities on Reddit of users of `r/The_Donald` and `r/fatpeoplehate`.

4.1 Operationalizing Reflection

Neo [26] describes **R**eflection as the emotional state making individuals receptive to radical narratives. To operationalize it, we compute a set of features quantifying the usage of toxic language, emotionality, anger, and anxiety in users' posts.

Language Toxicity (TOX). Users of radical communities are prone to use toxic language and engage in antisocial behavior, such as harassment, trolling, and cyberbullying [45]. Following Grover and Mark [14] suggestion that antisocial behavior can be captured through automated text analysis, we use the Perspective API [18] to measure language toxicity. We compute a user's toxicity (TOX) as the mean toxicity score of all their posts written before the ban on Reddit. Formally, the toxicity of a user i is $(\text{TOX})_i = 1/|P_i| \sum_{p \in P_i} \text{t}(c)$, where P_i is the set of posts of user i and $\text{t}(\cdot)$ is the toxicity score of a post.

Emotionality (EMO). We measure the intensity of emotions a user expresses as the average VADER score of their posts before the ban. VADER is a lexical rule-based sentiment analysis toolkit designed to identify sentiment in social media. It assigns a score to each word based on its positivity (negativity), and emotional charge. For example, the word "good" is less emotional than "awesome" even though both are positive. Formally, we define $(\text{EMO})_i$ of a user i as $1/|P_i| \sum_{p \in P_i} \text{VADER}(p)$. Where P_i is the set of posts of user i and $\text{VADER}(\cdot)$ is the VADER score of a post.

Anger (ANG) and Anxiety (ANX). Users of extreme online groups express anger and anxiety in their posts [11, 43]. We rely on the *Linguistic Inquiry and Word Count* (LIWC) dictionary, counting the proportion of anger (ANG)and anxiety (ANX)words a user i uses in their post written before the ban on Reddit.

4.2 Operationalizing Exploration

Neo [26] defines the **E**xploration as the period in which individuals make sense of the online information put forth by radical communities. We operationalize it by measuring (i) the diversity of interests and (ii) the engagement towards subreddits hosting discussions similar to those of `r/The_Donald` and `r/fatpeoplehate`.

Diversity of interests (DIV). To capture how frequently users of radical communities interact with other subreddits, we obtain the frequency of posts across subreddits and quantify how diverse user activity is with the *Gini coefficient*. The Gini coefficient measures the inequality among values of a frequency distribution. Thus, users contributing to a few subreddits have a low Gini coefficient, while those contributing to many have a higher score.

Engagement with radical communities (ENG). As measured by (DIV), users of radical communities can post in multiple subreddits. However, they may gravitate towards subreddits similar to their radical community, engaging with it outside a specific subreddit. To measure this engagement, we compute what proportion of the user's pre-ban activity is dedicated to posting in subreddits similar to the two focal subreddits (i.e., `r/The_Donald` and `r/fatpeoplehate`). We refer to this measure as (ENG), and we formalize it as $(\text{ENG})_i = \left(\sum_{s_{ij} \in S \setminus s_b} n_j \text{sim}(s_b, s_j) \right) / |P_i|$ where S is the set of all subreddits, P_i is the set of all posts made by users i, n_j is the number of posts made by user i on the j-th subreddit s_j, and s_b. $\text{sim}(s_b, s_j)$ is the similarity between the banned subreddit s_b (i.e., `r/The_Donald` and `r/fatpeoplehate`) and subreddit s_j (see appendix A. A high (ENG) score indicates that the user contributes primarily to subreddits close to their radical community.

4.3 Operationalizing Connection

The last factor we consider in users' radicalization is their **C**onnection to the radical online community, i.e., their influence on the user. We operationalize the concept by characterizing interactions with other community members.

Diversity of Interactions (DVI). Interactions with users exhibiting an *exclusive* interest in the radical communities might increase their sense of belonging to it [32]. We operationalize the magnitude of these interactions on a user i as the average Gini coefficient of users j they directly replied to, weighted by the number of comments exchanged between users i and j. Formally, we define diversity of interactions (DVI) as $\sum_{j \in \mathcal{N}_i} w_{ij}(\text{DIV})(j)/|\mathcal{N}_i|$. Where $\mathcal{N}_i$ is the set of those users that directly replied to user i, w_{ij} is the number of replies between user j and i. Finally, $(\text{DIV})(j)$ is computed as described in section 4.2

Influence of Seniority (SEN). Senior members of radical communities may exert pressure on users to conform to their views. We proxy the seniority on Reddit using the account's age (the number of days that elapsed from the first post on the community to the ban). We define (SEN) as the weighted average of the age differences between user i and the other users they directly replied to, i.e., $\frac{1}{|N_i|} \sum_{n \in N_i} w_{ij}[\alpha(i) - \alpha(n)]$. Where $\mathcal{N}_i$ is the set of those users that directly replied to user i, w_{ij} is the number of replies between user j and i. Finally, $\alpha(\cdot)$ is a function returning a user account's age.

Interactions with pre-ban migrated users. Interactions with other users who have joined the fringe platform might be before the ban (pre-ban migrated users) may create cross-platform ties that, following the ban, increase the odds of migrating to the fringe platform. We characterize these interactions in two ways:
*Active Interaction with pre-ban migrated users **(APB)**:* Active interactions are pairwise interactions between users of the radical community not yet posting on the fringe platform, with users already posting there and vice-versa. This exchange might be a user's first connection with the fringe platform through users already posting there. Therefore, we count the proportion of such dyadic interactions with pre-ban migrated users normalized by the user's dyadic interactions on Reddit before the ban.
*Passive Interaction with pre-ban migrated users **(PPB)**:* Users not yet posting on the fringe platform may post on threads where already migrated users also post. If a user posts on such a thread, we say that they interact *passively*. By posting on the same thread, users may passively consume the content written by pre-ban migrated users without directly engaging with them. Specifically, we calculate

the user's co-presence with pre-ban migrated users in threads by counting the threads with posts by pre-ban migrated users. We normalize this measure by the total number of threads the user participates in the pre-ban period, i.e., before the ban.

4.4 Other Variables

Language Coherence. Language Coherence is a concept proposed by Crossett and Spitaletta [5], Phadke et al. [29] expressing how well users align with the language of their community. Especially in discussions in radical communities, members might express their belonging to the community via their language [19, 38]. We measure language coherence using a language model capturing the linguistic state of the community. In particular, we fine-tune BERT on a dataset of posts written before the ban (we do not use these posts to measure language coherence). Then, given all pre-ban user posts on Reddit P_i, we estimate how unexpected the text is according to the language model fine-tuned on the community language. Specifically, we compute language coherence as the cross-entropy of all posts P_i for each user i given the language model.

Activity and Account Age. We add extra variables to eq. (2) and eq. (1) to control for self selection. Indeed, users who post on the fringe platform might be more active on the mainstream platform to begin with. Similarly, more senior users might be more motivated to post on the fringe platform as they develop a stronger attachment to the community. Therefore we use as control variables (i) *user activity* measured as the number of posts that the user made in the six months before the banning, and (ii) *user seniority* measured as the number of days elapsed from the first post on the community to the ban of the focal subreddit.

5 RESULTS

We show how radicalization factors affect users' migration with the regression analysis introduced in section 3. We perform separate analyses for `r/The_Donald` and `r/fatpeoplehate`. In both cases, we find that: (i) Reflection-related factors affects the first migration step, i.e., the decision to post on the fringe platform (FP) after the ban and (ii) Connection-related factors have a prominent role in the second migration step, i.e., the decision to be coactive on both the mainstream platform (MP) and fringe platform (FP). We conclude that individual factors primarily affect the first migration step, while coactivity is primarily affected by social factors. Table 1 and table 2 report the coefficients for the factors operationalizing Reflection, Exploration, and Connection for `r/The_Donald` and `r/fatpeoplehate`, respectively.

Table 1: Regression table for `r/The_Donald`. We show the parameters' estimates for the first and second stages of the Heckman regression.

	1st Step	2nd Step
	Selection Eq.	*Outcome Eq.*
(Intercept)	−1.02 (0.34)**	1.56 (0.70)*
Reflection		
Toxicity (TOX)	0.78 (0.15)***	−0.91 (0.37)*
Emotionality (EMO)	1.03 (0.20)***	0.42 (0.44)
Anger (ANG)	−1.79 (1.46)	−0.29 (1.03)
Anxiety (ANX)	−0.50 (1.12)	0.57 (0.97)
Exploration		
Diversification (DIV)	−2.93 (0.16)***	4.72 (1.15)***
Engagement (ENG)	1.25 (0.25)***	0.11 (0.51)
Connection		
Passive Int. (PPB)	0.23 (0.37)	1.56 (0.72)*
Active Int. (APB)	−0.01 (0.32)	7.56 (0.49)***
Neigh. Seniority (SEN)	0.37 (0.29)	4.22 (1.23)***
Neigh. Div. (DVI)	−0.09 (0.33)	2.44 (0.66)***
Controls		
Coherence	−5.22 (0.25)***	
Numb. Posts	2.39 (1.67)	0.55 (1.32)
Seniority	1.54 (0.63)*	0.54 (0.25)*
Rho (ρ)		0.28 (0.01)***
Sigma (σ)		0.32 (0.26)
AIC	8681.89	1433.90
Num. obs.	12053	2740

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

5.1 First Stage: Migration to Fringe Platforms

Reflection. Higher toxicity and emotionality on the MP are associated with more posts on the FP, as both (TOX) and (EMO) have positive coefficients for `r/The_Donald` ($\beta^{TD}_{\text{TOX}} = 0.78$, β^{TD}_{EMO}=1.03), and `r/fatpeoplehate` (β^{FPH}_{TOX}=1.12, β^{FPH}_{EMO}=0.73), respectively. In fig. 3, we show the marginal effects on the probability of posting on the FP. For example, we observe that, for `r/The_Donald`, a 20% increase in (TOX) increases the probability of becoming active on the FP by 8% (c.f., fig. 3a). An analysis of emotionality yields similar results. Reflection factors describing users' *individual* characteristics are significant predictors of the first migration decision. Interestingly, we find that the effect of (ANG) and (ANX) is not statistically significant.

Exploration. We find that Exploration factors, engagement (ENG) and diversity (DIV), are associated with the first migration decision. Specifically, the lower users' (DIV) is ($\beta^{\text{TD}}_{(\text{DIV})} = -2.93$, $\beta^{\text{FPH}}_{(\text{DIV})}$=−2.43), the more likely they are to post on the FP. For example, a 30% decrease in (DIV) increases the probability of posting on the FP by 18% (c.f., fig. 3b). Since (DIV) characterizes the heterogeneity of user interests, it acts as a 'pull factor' hindering the tendency to post on the FP. This result is in line with the survey study by Newell et al. [27]. Additionally, we find that users participating in other subreddits related to `r/The_Donald` or `r/fatpeoplehate` (*engagement*) are more likely to post on the fringe platform after the ban on the MP. In synthesis, Exploration, interpreted as the utilization of the mainstream platform, predicts the first migration step well.

Connection. We find that factors characterizing the connection to other community members are not statistically significant and, thus, do not correlate to the first migration decision. This is also shown in fig. 3c, the marginal effects of the features characterizing the social factors do not affect the probability of posting on FP.

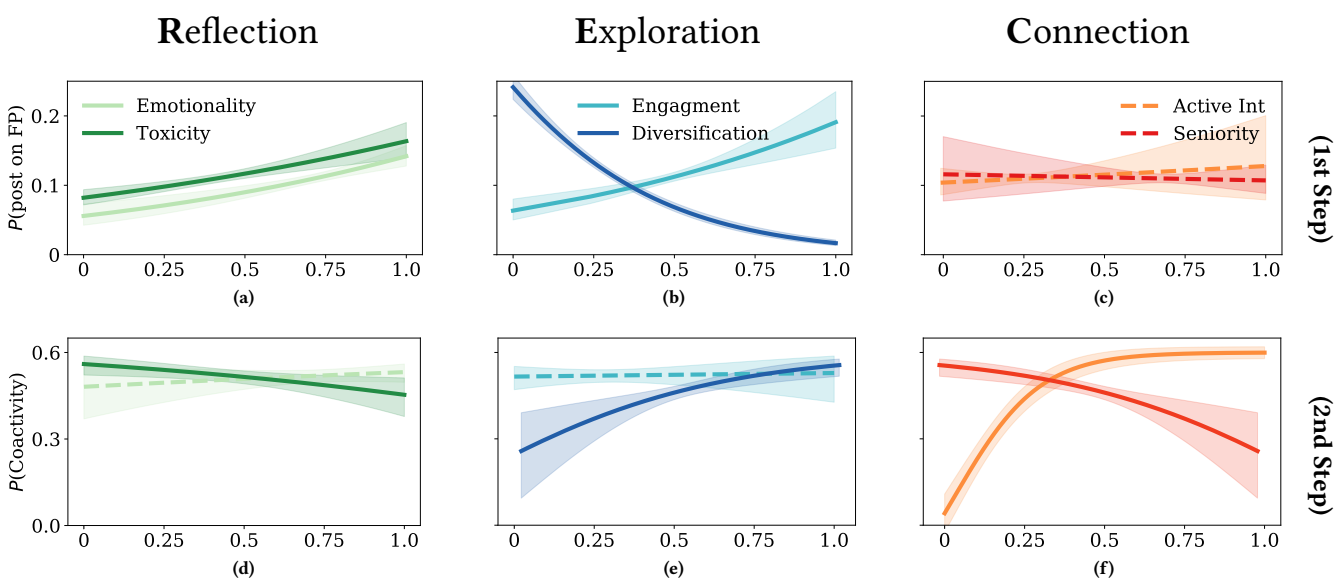

Figure 3: Marginal effects of REC factors for `r/The_Donald`. Solid lines represent factors with a significant effect. Dashed lines represent factors that do not have a significant effect. The shaded areas show the 95% confidence intervals. The marginals have been estimated assuming all other parameters are kept constant to their observed average values. *(top row)* Marginal effects for the first migration step. *(bottom row)* Marginal effects for the second migration step.

Table 2: Regression table for `r/fatpeoplehate`. We show the parameters' estimates for the first and second stages of the Heckman regression.

	1st Step	2nd Step
	Selection Eq.	*Outcome Eq.*
(Intercept)	−2.72 (0.57)**	2.06 (1.2)*
Reflection		
Toxicity (TOX)	1.12 (0.14)***	−1.06 (0.64)
Emotionality (EMO)	0.73 (0.20)**	0.34 (0.52)
Anger (ANG)	0.53 (1.82)	−0.12 (0.41)
Anxiety (ANX)	−0.05 (1.30)	0.41 (0.76)
Exploration		
Diversification (DIV)	−2.43 (0.21)***	3.52 (0.54)***
Engagement (ENG)	1.48 (0.52)**	0.17 (0.58)
Connection		
Passive Int. (PPB)	0.36 (0.23)	1.96 (0.46)*
Active Int. (APB)	−0.05 (0.41)	5.56 (0.88)***
Neigh. Seniority (SEN)	1.17 (1.09)	3.27 (1.01)**
Neigh. Div. (DVI)	−0.22 (0.19)	1.92 (0.66)**
Controls		
Coherence	−4.87 (0.45)**	
Numb. Posts	3.21 (2.72)	1.88 (1.52)
Seniority	0.07 (0.39)	0.51 (0.27)*
Rho (ρ)		0.36 (0.02)***
Sigma (σ)		0.44 (0.12)*
AIC	7953.34	1681.89
Num. obs.	8168	916

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

5.2 Second stage: Coactivity across Platforms

Reflection. Reflection factors determine users' coactivity to a minor extent. Toxicity (TOX) is the only significant reflection factor in the case of `r/The_Donald`, users that are less toxic in the mainstream platform have a higher chance of being coactive after the ban ($\beta^{TD}_{TOX} = -0.91$). For `r/fatpeoplehate`, no reflection factor is statistically significant. In fig. 3d, we show that both (TOX) and (EMO) do not increase the probability of being coactive.

Exploration. *Diversification* affect users' coactivity as in the first migration step. The lower the users' (DIV), the more they behave as coactive users ($\beta^{TD}_{DIV} = 4.72$, $\beta^{FPH}_{DIV} = 3.52$). For instance, in fig. 3e, we show for `r/The_Donald` that a 30% decrease in (DIV) increases by 22% the probability of coactivity. Unlike the first migration step, (ENG) does not affect coactivity.

Connection. All Connection's factors become statistically significant and strongly affect users' coactivity. In particular, from table 1 and table 2, we observe that the more users directly interact with pre-ban migrated users (APB), the less they tend to stay coactive. Further, we find that users interacting with more recent members are more likely to be active on both platforms after the ban (SEN). In fig. 3f, we observe that directly interacting with pre-migrated users increases the probability of coactivity by up to 70% in the case of `r/The_Donald` subreddit.

6 DISCUSSION

This article studies the post-ban migration of users of radical communities from mainstream to fringe platforms. Migrating users of radical communities, i.e., those who become active on fringe platforms, either abandon the mainstream platform or remain coactive

on both. Previous work has shown that the migration decision following radical community bans has negative externalities. Namely (i) the creation of more toxic communities elsewhere [16] (if many users decide to migrate); and (ii) the spillover of toxic behaviour from the fringe onto the mainstream platform (if many users decide to be coactive).

To understand the factors associated with each decision (migration and coactivity), we conduct a two-step regression analysis. The first step estimates the probability of users posting on the fringe platform (migration decision). The second step estimates the propensity of users to be coactive (coactivity decision). Specifically, we examine two subreddits, `r/The_Donald` and `r/fatpeoplehate`, associated with toxic behavior.

Our analysis reveals how factors of the RECRO radicalization framework relate to users' migration decisions. Our findings are threefold. First, individuals' online behavior, described by Reflection, capturing the needs and vulnerabilities of users, is linked with the decision to post on the fringe platform. For example, users that exhibit higher toxicity (TOX) or emotionality (EMO) migrated more often. Second, interactions with the social environment described by Connection relate *primarily* to the decision to be coactive on both platforms. Surprisingly, we find that interactions with pre-ban migrated users (APB)/(PPB) increase the propensity to post on the fringe platform. Instead, users with more such interactions are more likely to remain coactive. Third, the diversity of users' interests, captured by Exploration, hinders them from abandoning the mainstream platform altogether and increases their propensity to remain coactive. This is in accordance with previous findings suggesting that the diversity of content on mainstream platforms like Reddit is a pull factor that limits full migration [27]. These findings suggest that the decision to engage with the new platform is linked to individual motives, whereas social factors are associated with coactivity.

Implications. Our work has two implications. First, it can help moderators to predict users' reactions to community bans. Thus, paving the way for moderators to take more informed decisions on banning. For instance, platforms like Reddit could estimate how users will react before carrying out community bans. Second, our findings contribute to a growing literature on understanding online radicalization. Fringe platforms have been tightly linked with terrorist attacks and extremist ideologies [46], and therefore, studying what makes users migrate to fringe platforms advances our understanding of radicalization.

Limitations. Our classification of users according to their posting activity on fringe platforms may be inaccurate and thus bias our results. For example, some users may change their username or only read posts on the fringe platform, and we would erroneously classify them as Reddit-only when they are participating or consuming content from the FP. Our result remains unbiased, however, if there is no systematic difference between users keeping their usernames across platforms and those changing them.

REFERENCES

[1] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Understanding the Effect of Deplatforming on Social Networks. In *13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 187–195. https://doi.org/10.1145/3447535.3462637

[2] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839.

[3] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[4] Ben Collins and Brandy Zadrozny. 2020. Facebook bans QAnon across its platforms. https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339.

[5] Chuck Crossett and Jason Spitaletta. 2010. Radicalization: Relevant psychological and sociological concepts. *The John Hopkins University* (2010).

[6] Anja Dalgaard-Nielsen. 2010. Violent radicalization in Europe: What we know and what we do not know. *Studies in conflict & terrorism* 33, 9 (2010), 797–814.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Doggoes. 2020. 'I hope if you came from T_D you reserved your reddit username even if you don't plan to useit'. https://thedonald.win

[9] Claudia I Flores-Saviaga, Brian C Keegan, and Saiph Savage. 2018. Mobilizing the trump train: Understanding collective action in a political trolling community. In *Twelfth International AAAI Conference on Web and Social Media*.

[10] Deen Freelon, Alice Marwick, and Daniel Kreiss. 2020. False equivalencies: Online activism from left to right. *Science* 369, 6508 (2020), 1197–1201.

[11] Jörg Friedrichs, Niklas Stoehr, and Giuliano Formisano. 2022. Fear-anger contests: Governmental and populist politics of emotion. *Online Social Networks and Media* 32 (2022), 100240. https://doi.org/10.1016/j.osnem.2022.100240

[12] Google. 2022. YouTube Community Guidelines enforcement. https://transparencyreport.google.com/youtube-policy/removals.

[13] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.

[14] Ted Grover and Gloria Mark. 2019. Detecting potential warning behaviors of ideological radicalization in an alt-right subreddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 193–204.

[15] James J. Heckman. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47, 1 (Jan. 1979), 153. https://doi.org/10.2307/1912352

[16] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. 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), 1–24.

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

[18] Jigsaw. 2022. Perspective API. https://perspectiveapi.com/.

[19] Daniel Koehler. 2014. The radical online: Individual radicalization processes and the role of the Internet. *Journal for Deradicalization* 1 (2014), 116–134.

[20] Arie W Kruglanski, Michele J Gelfand, Jocelyn J Bélanger, Anna Sheveland, Malkanthi Hetiarachchi, and Rohan Gunaratna. 2014. The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology* 35 (2014), 69–93.

[21] Walter Laquer. 1998. *Origins of terrorism: Psychologies, ideologies, theologies, states of mind.* Woodrow Wilson Center Press.

[22] Matthew N Lyons. 2017. Ctrl-alt-delete: The origins and ideology of the alternative right. *Political Research Associates* 20 (2017).

[23] Alice E Marwick and Robyn Caplan. 2018. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies* 18, 4 (2018), 543–559.

[24] Reid McIlroy-Young and Ashton Anderson. 2019. From "welcome new gabbers" to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aaai conference on web and social media*, Vol. 13. 651–654.

[25] Amin Mekacher and Antonis Papasavva. 2022. " I Can't Keep It Up." A Dataset from the Defunct Voat. co News Aggregator. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 1302–1311.

[26] Loo Seng Neo. 2019. An Internet-mediated pathway for online radicalisation: RECRO. In *Violent Extremism: Breakthroughs in Research and Practice*. IGI Global, 62–89.

[27] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on reddit during a period of community unrest. In *Tenth International AAAI Conference on Web and Social Media*.

[28] Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. Soros, child sacrifices, and 5G: understanding the spread of conspiracy theories on web communities. *arXiv preprint*

arXiv:2111.02187 (2021).
[29] Shruti Phadke, Mattia Samory, and Tanushree Mitra. 2022. Pathways through Conspiracy: The Evolution of Conspiracy Radicalization through Engagement in Online Conspiracy Discussions. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 770–781.
[30] Patrick Puhani. 2000. The Heckman correction for sample selection and its critique. *Journal of economic surveys* 14, 1 (2000), 53–68.
[31] Reddit. 2020. Update to our content policy. https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/.
[32] Justin Reedy, John Gastil, and Michael Gabbay. 2013. Terrorism and small groups: An analytical framework for group disruption. *Small group research* 44, 6 (2013), 599–626.
[33] Diana Rieger, Anna Sophie Kümpel, Maximilian Wich, Toni Kiening, and Georg Groh. 2021. Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and Reddit. *Social Media+ Society* 7, 4 (2021), 20563051211052906.
[34] Giuseppe Russo, Christoph Gote, Laurence Brandenberger, Sophia Schlosser, and Frank Schweitzer. 2022. Disentangling Active and Passive Cosponsorship in the US Congress. *arXiv preprint arXiv:2205.09674* (2022).
[35] Giuseppe Russo, Nora Hollenstein, Claudiu Cristian Musat, and Ce Zhang. 2020. Control, Generate, Augment: A Scalable Framework for Multi-Attribute Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 351–366. https://doi.org/10.18653/v1/2020.findings-emnlp.33
[36] Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2022. Spillover of Antisocial Behavior from Fringe Platforms: The Unintended Consequences of Community Banning. *arXiv preprint arXiv:2209.09803* (2022).
[37] Mattia Samory and Tanushree Mitra. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[38] Ryan Scrivens, Garth Davies, and Richard Frank. 2018. Searching for signs of extremism on the web: an introduction to Sentiment-based Identification of Radical Authors. *Behavioral sciences of terrorism and political aggression* 10, 1 (2018), 39–59.
[39] Andrea Sipka, Aniko Hannak, and Aleksandra Urman. 2022. Comparing the Language of QAnon-related content on Parler, Gab, and Twitter. In *14th ACM Web Science Conference 2022*. 411–421.
[40] Niklas Stoehr, Lucas Torroba Hennigen, Josef Valvoda, Robert West, Ryan Cotterell, and Aaron Schein. 2022. An Ordinal Latent Variable Model of Conflict Intensity. *arXiv preprint arXiv:2210.03971* (2022).
[41] Nathalie Van Raemdonck. 2019. The echo chamber of anti-vaccination conspiracies: mechanisms of radicalization on Facebook and Reddit. *Institute for Policy, Advocacy and Governance (IPAG) Knowledge Series, Forthcoming* (2019).
[42] Isaac Waller and Ashton Anderson. 2021. Quantifying social organization and political polarization in online platforms. *Nature* 600, 7888 (2021), 264–268.
[43] Dag Wollebæk, Rune Karlsen, Kari Steen-Johnsen, and Bernard Enjolras. 2019. Anger, fear, and echo chambers: The emotional basis for online behavior. *Social Media+ Society* 5, 2 (2019), 2056305119829859.
[44] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. 188–202.
[45] Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. In *12th ACM Conference on Web Science*. 125–134.
[46] Jing Zeng and Mike S Schäfer. 2021. Conceptualizing "dark platforms". Covid-19-related conspiracy theories on 8kun and Gab. *Digital Journalism* 9, 9 (2021), 1321–1343.
[47] Ethan Zuckerman and Chand Rajendra-Nicolucci. 2021. Deplatforming Our Way to the Alt-Tech Ecosystem. *Knight First Amendment Institute at Columbia University, January* 11 (2021).

A APPENDIX

A.1 Subreddit Similarity

We define a similarity between all subreddits and the two focal subreddits (`r/The_Donald` and `r/fat peoplehate`) considering their polar opposites (`r/HillaryClinton` and `r/fatlogic`, respectively). Given a focal subreddit s_i, we define a subreddit s_j to be *relevant* for s_i if at least ten users of either the focal subreddit s_i or its polar opposite $s_{\hat{i}}$ posted at least five times on s_j before the ban. We then construct a weighted graph for a focal subreddit s_i and its polar opposite $s_{\hat{i}}$ (e.g., `r/fatpeoplehate-r/fatlogic`). The nodes of the graphs consist of (i) s_i and $s_{\hat{i}}$, and (ii) all relevant subreddits for the two polar opposites. We draw a weighted edge between two nodes if the corresponding subreddits share at least five active users. The weight corresponds to the number of users shared. As a next step, we train the Node2Vec [13] algorithm on the graphs built for the pairs `r/The_Donald-r/HillaryClinton` and `r/fatpeoplehate-r/fatlogic`. As a result, we obtain embeddings specifically catered towards finding subreddits hosting discussions similar to the focal subreddits. We use the cosine similarity to map subreddits' similarity on a scale from -1 to +1, where +1 represents the higher similarity to the focal subreddits. To validate our similarity scale, we calculate Spearman's rank-order correlation between the 1,000 subreddits most similar to `r/The_Donald`, and `r/fatpeoplehate` and the ranking of publicly available subreddit embeddings by Waller and Anderson [42]. The embeddings from Waller and Anderson [42] are not explicitly trained towards finding similarities between specific communities, but they provide a general measure of subreddit similarity. We find a correlation of 0.64 with $p < 0.05$, indicating that our scale successfully measures similarity to `r/The_Donald` and `r/fatpeoplehate`. We manually inspect the top 50 subreddits on the similarity scale and confirm that they host discussions similar to `r/The_Donald` and `r/fatpeoplehate`, respectively.

A.2 Predicting Migration Steps

Table 3: F1-Scores for the different classification models. The first column reports the classification scores on the three labels RO, CA, and FM. The second and third column report the classification scores of the first (RO ↔ MG) and second step (CA ↔ FM) respectively of the 2STEP classifier.

	F1-Score (All)	F1-Score (1st step)	F1-Score (2nd Step)
SVM	0.44	~	~
Random Forest	0.52	~	~
XGB	0.65	~	~
AdaBoost	0.41	~	~
2STEP (Our)	**0.74**	0.84	0.88

In the analysis presented in section 3 and section 5, we show how reflection, exploration, and connection factors affect user migration. Here, we investigate if and how well we can predict users' migration decisions in reaction to a ban. To answer these questions, we define a hierarchical classification model. This classifier first distinguishes users who post on a fringe platform from those who will not (first decision). Subsequently, a second classifier predicts if they become co-active or not. This second classifier is trained with those users that, according to the first classifier, will post on the fringe platform. We use for both the first and second classifier a Gradient Boosting Classifier.

To correct the high imbalance that primarily affects the first classifier, we downsample the users that did not post on the fringe platform as they are ten times more common than those that post on the fringe. We evaluate our hierarchical classifier against other models using multiple metrics testing prediction capacities. All

the experiments have been conducted by running a five-fold cross-validation using a 60%-20%-20% training-validation-test split.

Our results show that our model outperforms other baselines that do not consider the two-step structure of online migration. Specifically, we compare the performance of our classifier against baselines that directly predict the final user decision in a single step (Reddit-Only, Coactive, or Fully-Migrated). Unlike our classifier, these baselines predict users' migration decisions in a single step. In our experiments, the 2STEP classifier outperforms all baselines by a maximum of 12% using f1-score accuracy. These results provide solid evidence that our two-step characterization of the migration process is a valid assumption. We report the results of our classification in table 3. Such results provide additional support to the analysis that we conducted in section 3.2 Additionally, recent advances in natural language processing [7, 35] and graph learning [34] can help in improving the prediction performance of this prediction task.