Stranger Danger! Cross-Community Interactions with Fringe Users Increase the Growth of Fringe Communities on Reddit

Giuseppe Russo, Manoel Horta Ribeiro, Robert West
EPFL
giuseppe.russo@epfl.ch, manoel.hortaribeiro@epfl.ch, robert.west@epfl.ch

Abstract

Fringe communities promoting conspiracy theories and extremist ideologies have thrived on mainstream platforms, raising questions about the mechanisms driving their growth. Here, we hypothesize and study a possible mechanism: new members may be recruited through *fringe-interactions*: the exchange of comments between members and non-members of fringe communities. We apply text-based causal inference techniques to study the impact of fringe-interactions on the growth of three prominent fringe communities on Reddit: r/Incel, r/GenderCritical, and r/The_Donald. Our results indicate that fringe-interactions attract new members to fringe communities. Users who receive these interactions are up to 4.2 percentage points (*pp*) more likely to join fringe communities than similar, matched users who do not.This effect is influenced by 1) the characteristics of communities where the interaction happens (e.g., left vs. right-leaning communities) and 2) the language used in the interactions. Interactions using toxic language have a 5*pp* higher chance of attracting newcomers to fringe communities than non-toxic interactions. We find no effect when repeating this analysis by replacing fringe (r/Incel, r/GenderCritical, and r/The_Donald) with non-fringe communities (r/climatechange, r/NBA, r/leagueoflegends), suggesting this growth mechanism is specific to fringe communities. Overall, our findings suggest that curtailing fringe-interactions may reduce the growth of fringe communities on mainstream platforms.

1 Introduction

Mainstream platforms enacted various moderation policies to curtail fringe communities due to their association with real-world violence and online harassment (Collins and Zadrozny 2020). For instance, r/The_Donald, a Reddit community that was key in planning the 2021 US Capitol invasion (Washington Post 2020), was extensively sanctioned by Reddit, having its visibility reduced and being removed from the main feed. However, even amidst sanctions, fringe communities still flourish and attract new members (Trujillo and Cresci 2022; Horta Ribeiro et al. 2021), e.g., even after the sanctions, r/The_Donald remained one of the most active communities on Reddit, with over 790,000 users before being banned. This observation suggests that visibility through platform affordances (e.g., Reddit's front page) may not be the sole mechanism driving the attraction of new members.

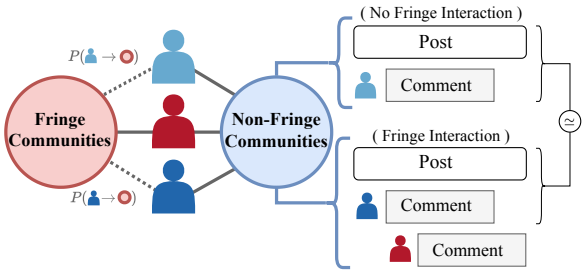

Figure 1: We study whether interacting with fringe users ( ) causes non-fringe users ( ) to join fringe communities. We match users that involuntarily received comments from fringe users (; "treatment") with users who did not, but had similar chances to do so (; "control;" ≃). Then, we estimate the causal effect of fringe interactions by comparing the fraction of users that went on to join the fringe community in the treatment ($P(\rightarrow \bullet)$) and in the control group ($P(\rightarrow \bullet)$).

Present work. Here, we consider another potential mechanism for the growth of fringe communities on mainstream platforms: *fringe-interactions*, defined as an exchange of comments between users of fringe communities and other users active on the same platform (e.g., Reddit) but not in the fringe community (e.g., r/The_Donald) at the time of the interaction. We hypothesize that fringe-interactions boost the visibility of fringe communities by exposing non-fringe users to ideas or concepts associated with the fringe community via either the profile of fringe users or the content of the interactions.

We center our analysis around three research questions:

RQ1: Do fringe-interactions drive newcomers toward fringe communities? If so, is this specific to fringe communities?

RQ2: In which communities are fringe-interactions more successful in attracting newcomers?

RQ3: What linguistic traits characterize fringe-interactions that successfully attract newcomers?

To answer these questions, we study three prominent fringe communities on Reddit (r/Incel, r/GenderCritical, and r/The_Donald) using the quasi-experimental setup illustrated in fig. 1.

First (**RQ1**), we use regression analysis to measure the probability of users joining fringe subreddits after interacting with fringe users. We then compare this probability to that of users who did *not* interact with fringe users. Via balanced risk set matching (Rosenbaum and Rubin 1983), we ensure that these two groups of users have comparable propensities to interact with fringe users. To understand if this mechanism is specific to fringe communities, we repeat the analysis, considering interactions between members of *non-fringe* subreddits (r/climatechange, r/NBA, r/leagueoflegends) and other users not yet active in these subreddits. Second (**RQ2**), we analyze how the effect of fringe-interactions varies depending on the characteristics of the community (e.g., political leaning) where the fringe-interaction occurred. Third (**RQ3**), we characterize the content of fringe-interactions that successfully attract newcomers across a variety of linguistic traits (e.g., toxicity).

Findings. Our analysis shows that fringe-interactions increase the likelihood of a user posting on a fringe community by up to 4.2 percentage points (*pp*) (**RQ1**). We do not find evidence that interactions with non-fringe users (e.g., members of r/climatechange) drive newcomers towards non-fringe communities (r/climatechange itself), suggesting that the growth through interactions is specific to fringe communities.

Additionally, we find that the strength of the effect of fringe-interactions varies depending on the characteristics (political orientation, gender, and age) of the subreddits where the interactions occur (**RQ2**). For instance, users interacting with fringe users in right-leaning subreddits are 15.6 *pp* more likely to join fringe communities than those active in the same subreddit who have not interacted with fringe users. Finally, our analysis indicates that interactions containing toxic language increase the chances of attracting newcomers to fringe communities by 5*pp* on average (**RQ3**).

Implications. Mainstream online platforms have attempted to curtail fringe communities by banning them or reducing their visibility. However, neither of these interventions is a silver bullet, as they are remarkably resilient (Horta Ribeiro et al. 2021; Trujillo and Cresci 2022). The growth mechanism studied here shows that fringe-interactions are another target for stakeholders. For instance, platforms could reduce the visibility of fringe community members *outside* of fringe communities or proactively target users who receive messages from fringe community members with information that may reduce their chance to join fringe communities.

2 Related Work

Why do people join online communities? Scholars have extensively studied user motivations for joining online communities (Ridings and Gefen 2004; Ren et al. 2012). Past research indicates that users deliberately seek communities that align with their social identities (Ammari and Schoenebeck 2015; Lingel, Naaman, and Boyd 2014), and has highlighted the importance of content quality (Lu, Phang, and Yu 2011; Zhang et al. 2017), effective moderation (Lampe and Johnston 2005), and meta-characteristics like size, activity levels, and network structures (Hwang and Foote 2021). Further, Backstrom et al. (2008) has found that interpersonal interactions with other users draw individuals to online communities. In the case of fringe communities, research on why users participate in these communities focused on users' psychological characteristics (Schmid 2013), finding that hopelessness, sadness, and anxiety are the primary psychological drivers of participation (Caren, Jowers, and Gaby 2012; Lauckner and Hsieh 2013).

Fringe communities. We briefly describe the fringe communities considered in this study. r/The_Donald, created in June 2015 to support Donald Trump's presidential campaign, became associated with the "alt-right" movement, hosting discussions involving racism, sexism, and Islamophobia. It also spread conspiracy theories and was mobilized for "political trolling" (Lyons 2017; Paudel et al. 2021). r/GenderCritical, created in September 2013, hosted the trans-exclusionary radical feminists (TERFs) community, known for doxing and harassing trans women (Kaitlyn 2020; Williams 2020). r/Incel, created in August 2013, was a community of self-denominated "involuntary celibates" adhering to "The Black Pill," the belief that unattractive men are doomed to romantic loneliness and unhappiness (Ribeiro et al. 2021). Since their inception, the Incels community has been closely related to terrorist attacks and the production of misogynistic content online (Jaki et al. 2019; Hoffman, Ware, and Shapiro 2020).

Interventions against fringe communities. r/Incel, r/The_Donald, and r/GenderCritical have all been banned due to breaching Reddit's guidelines. Previous work has studied the effectiveness of these bans, showing that banned users reduce their activity on mainstream platforms (Chandrasekharan et al. 2017; Jhaver et al. 2021). However, part of the banned community migrates to other fringe platforms (Monti et al. 2023; Russo et al. 2023a) causing possible spillovers of antisocial behavior back onto the mainstream platform (Russo et al. 2023b; Schmitz, Muric, and Burghardt 2022). Other works have examined softer interventions like limiting the visibility of fringe communities and warning new visitors about the potential issues with content posted in these communities (Chandrasekharan et al. 2022; Trujillo and Cresci 2022), finding that these measures reduce the number of newcomers to some extent. Overall, this literature suggests that interventions against fringe communities are no silver bullet, as they are highly adversarial and have highly engaged members.

Relation between present and prior work. Here, we explore fringe-interactions, a potential mechanism that may contribute to the growth of fringe communities. The existence of this mechanism is backed by previous research highlighting the importance of interpersonal interaction in drawing users to new online communities (Backstrom et al. 2008; Corso, Russo, and Pierri 2024). Given how challenging it is to reduce the influence of fringe communities in our online ecosystem (Horta Ribeiro et al. 2021; Russo, Stoehr,

	Comments	Users	Selected Users	Period Considered	Treated	Potential Controls
Fringe subreddits						
r/Incels	2,041,313	177,095	38,145	06/2016 – 06/2017	86,556	5,275,321
r/GenderCritical	1,629,169	48,243	9,624	01/2019 – 01/2020	54,142	2,250,641
r/The_Donald	12,387,349	482,796	72,371	05/2018 – 05/2019	97,823	6,481,223
Non-Fringe subreddits						
r/climatechange	632,674	93,256	19,374	01/2019 – 01/2020	81,307	3,912,866
r/NBA	1,483,631	468,321	67,481	01/2018 – 01/2019	72,321	2,973,364
r/leagueoflegends	3,573,974	694,522	71,944	01/2019 – 01/2020	95,763	6,514,447

Table 1: Data collection summary — For the three fringe subreddits considered in this paper, we show the number of comments, users, and users obtained (columns 2–4), as well as the treated and potential control units associated with these users' fringe-interactions (columns 6–7) across one-year periods (column 5). We also show the same statistics for non-fringe subreddits that we use to repeat the experiment.

and Ribeiro 2023), we argue that understanding why these communities thrive is key to increasing the arsenal of interventions available to curtail their growth.

3 Material and Methods

Data

Reddit data. To answer our research questions, we selected three prominent fringe subreddits r/Incels, r/GenderCritical, and r/The_Donald (see Section 2). Using the Pushshift API (Baumgartner et al. 2020), we retrieve all comments made in these three fringe subreddits from their creation to their banning from Reddit. We consider as part of r/Incels, r/GenderCritical, or r/The_Donald only those users that post more than five comments in one of these three subreddits. This is similar to what has been done in previous research (Kumar et al. 2018; Samory and Mitra 2018). In cases where a user exceeds the threshold in multiple fringe subreddits, we consider the user as a member of the subreddit where they have posted the most; this prevents fringe users from being considered multiple times across different fringe subreddits. We collect roughly 15M posts from 700K users from these three fringe subreddits.

To understand if the growth through interactions is specific to fringe communities, we repeat our analyses, replacing fringe with *non*-fringe subreddits (r/climatechange, r/NBA, and r/leagueoflegends). Therefore, we collect all posts made on three non-fringe subreddits, r/climatechange, r/NBA, and r/leagueoflegends, chosen to represent a diverse set of communities. [1] We collect roughly 5M posts from 1M users from these three non-fringe subreddits. We provide statistics for these collection processes in Section 1.

Treatment and control groups. Our study centers on understanding how interactions with fringe users influence the attraction of newcomers to fringe subreddits. To achieve this, we compare the likelihood of users joining a fringe subreddit after a fringe-interaction (*treatment group*) with the likelihood of other users joining the same fringe subreddit without any interaction (*control group*). To identify users in the treatment group, we gather all fringe-interactions made by users of fringe subreddits on subreddits other than fringe ones. In other words, we look for interactions occurring in subreddits different from r/Incels, r/GenderCritical, or r/The_Donald or related to them (e.g., Incels2).

A fringe-interaction consists of a comment c_{fringe} made by a fringe user u_{fringe} on a non-fringe subreddit $s_{\text{non-fringe}}$ in response to a comment $c_{\text{non-fringe}}$ made by a non-fringe user $u_{\text{non-fringe}}$ on the same non-fringe subreddit $s_{\text{non-fringe}}$. We include non-fringe users in our treatment group if the following conditions are met:

1. the non-fringe user $u_{\text{non-fringe}}$ did not post in the fringe subreddit (e.g., r/Incels) associated with the fringe user u_{fringe} prior to the interaction.
2. u_{fringe} has never interacted before with $u_{\text{non-fringe}}$. In other words, they never exchanged a comment
3. The comment c_{fringe}, authored by u_{fringe}, occurred within a week from the comment $c_{\text{non-fringe}}$ posted by $u_{\text{non-fringe}}$.

Finally, following previous research from Phadke, Samory, and Mitra (2022), we collect our data only from subreddits that received at least five contributions from users of fringe subreddits and consider fringe-interactions that happened within one year. To define a control group, we collect all comments made in the same weeks and subreddits as the ones in the treatment group where no fringe-interaction happened. We provide these statics in Section 1

Outcome. Given users that receive (treatment) or not (control) an interaction from a user active on a fringe subreddit (e.g., r/Incels), our outcome variable is the number of users joining the same fringe subreddit in the weeks following the interaction. Specifically, we analyze users' post and comment activity in the eight weeks following the date of comments associated with users in our treatment/control groups. We consider that a user "joined" a fringe subreddit if they post or comment at least once in that subreddit after the interaction. We operationalize these covariates following the same approach when repeating our analysis considering

[1] Respectively, a political movement, a community centered around an "offline" event, and a community centered around an online video game.

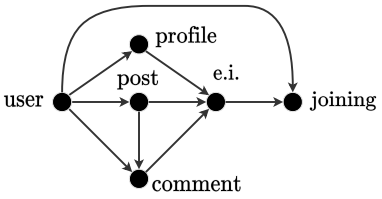

Figure 2: Causal diagram. Given a comment $c_{\text{non-fringe}}$ made by a non-fringe user $u_{non-fringe}$, we assume that a fringe-interaction depends exclusively on (1) the *profile* of the non-fringe who posted the comment $c_{\text{non-fringe}}$, (2) on the *post* where the comment was made, and (3) on the content of the *comment* itself. We control for these factors to estimate the effect of a fringe-interaction in subsequent participation in a fringe community (*fringe int.* $\rightarrow$ *joining*).

the interactions of non-fringe users from r/climatechange, r/NBA, and r/leagueoflegends.

Identification

Our analysis uses observational data to mimic a hypothetical experiment. Suppose Reddit implemented a filter where whenever a user who is active in a fringe community replies to someone in a non-fringe community, the comments of the fringe user do not get shown. Then, for a select, randomly assigned treatment group, this filter gets removed. Conversely, nothing changes for the remaining control users; the filter remains as is. We are interested in the difference in the rate at which treatment vs. control group users join the fringe community.

Since this experiment is not feasible, we use observational data as a substitute. Our treatment group consists of users who interacted with a fringe user from either r/Incels, r/GenderCritical, or r/The_Donald. On the other hand, our control group consists of users who were equally likely to interact with a fringe user but did not. To identify the effect of fringe-interactions on non-fringe users, we formalize our assumptions in the causal graph in fig. 2. We assume that given a comment $c_{\text{non-fringe}}$ (*comment* in the causal graph) whether it receives a fringe-interaction depends on (1) the profile of the user $u_{\text{non-fringe}}$ (*profile* in the causal graph) who posted the comment $c_{\text{non-fringe}}$, (2) the content of the comment $c_{\text{non-fringe}}$ (*comment*), and (3) the content of the post p (*post*) that corresponds to the Reddit submission the comment $c_{\text{non-fringe}}$ refers to.

Under these assumptions, we can control for confounders by blocking the biasing paths (Glymour, Pearl, and Jewell 2016) to isolate the causal effect of fringe-interactions. We do so by controlling the user profiles, the content, and the structure of comments and posts, respectively. Indeed, different comments/posts may draw attention differently, stimulating possible interactions from fringe users. However, other confounders may affect the likelihood of non-fringe users to join fringe subreddits *regardless of fringe-interactions*, e.g., being a young male with relationship issues makes someone more likely to join r/Incels, we address this issue in detail in Section 5.

Operationalization

Below, we describe how we operationalize the confounding variables in the causal diagram in Figure 2. Note that these covariates are computed for a triplet $\langle \text{comment}, \text{post}, \text{profile} \rangle$, where, in the treatment group, the user who made the comment interacted with a fringe user, and in the control group, they did not.

Comment. When comparing comments that received fringe-interactions with those that did not, we want the content of these comments to be semantically similar. In a recent comparison of text-level adjustment strategies, Weld et al. (2022) found that transformer-based representations outperformed other text representations, and thus, we use a fine-tuned version of the BERT architecture (Devlin et al. 2018) to obtain a representation for each comment in our dataset. We describe this in detail in Section 3.

Post. We also want the posts associated with comments (potentially) receiving fringe-interaction to be similar. This is partially addressed by matching comments within similar subreddits, but in addition, we also consider, for each post, the number of (1) direct replies that a post received, (2) unique users that directly replied to the post, (3) comments in the post, and (4) unique user commenting in the post. Importantly, these variables are computed prior to the creation of the comment of interest.

Profile. Fringe users may choose to interact with a non-fringe user (via a comment) because of the profile of the non-fringe user that posted the comment. To control for the profile, we characterize each non-fringe user using their activity and its relatedness to the fringe subreddit (e.g., r/Incels). Specifically, we operationalize (1) *user activity* as the total number of posts made in the 8 weeks before the comment, and (2) the *fringe score* as the proportion of the user's Reddit activity dedicated to discussion related to the fringe subreddit. To calculate the *fringe score*, we follow Phadke, Samory, and Mitra (2022): for a non-fringe user u that write N_u posts in a set S_u of subreddits the fringe score f_u is

$$f_u = \frac{\sum_{s \in S_u} n_s \text{sim}(s_{\text{fringe}}, s)}{N_u}, \tag{1}$$

where n_s is the number of comments made on the subreddit s, $\text{sim}(s_{\text{fringe}}, s)$ is the cosine similarity between the embeddings (from Waller and Anderson (2021)) of the fringe subreddit s_{fringe}, (e.g., r/Incels) and a subreddit s.

Estimation

Having operationalized the treatment, the outcome, and the control variables of interest, we then estimate the causal effect of fringe-interactions. Our approach is two-fold: we match comment-user pairs ($c_{non-fringe}$, $u_{non-fringe}$) that received fringe-interactions with similar pairs that did not. We then conduct a regression analysis in the subset of matched comments posted by non-fringe users. This approach allows us to obtain robust results by combining the strengths of both approaches (Gelman 2014).

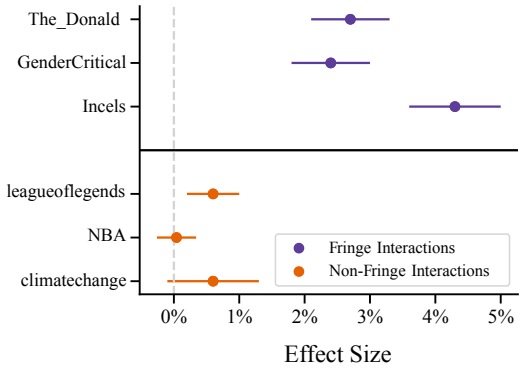

Figure 3: The effect of fringe-interactions on joining fringe communities— Differential increase in the probability of joining fringe subreddits after interacting with fringe users. Significant effect observed for r/Incels, r/GenderCritical, and r/The_Donald. We repeat the experiment considering interactions with users of non-fringe communities finding not significant or very small effects. Significance level at 0.05

Balanced Risk Set Matching. In our setting, users in the treatment group receive the treatment (fringe-interactions) at different times. This is because fringe users reply to comments of non-fringe users at different times. When a treatment is given at various times, it is important to form matched pairs in which subjects are similar prior to treatment but avoid matching on events subsequent to treatment. This is done using risk-set matching, in which a newly treated individual at time t is matched to one control not yet treated at time t based on covariate information describing subjects prior to time t. The covariates we must control for with the matching (see Figure 2) vary with time, e.g., a non-fringe user have a similar profile to another non-fringe user only in time t_1, not t_2.

We address this issue by performing Balanced Risk Set Matching (Rosenbaum and Rubin 1983), a procedure that consists of (1) creating "risk sets," groups of individuals that did not receive the treatment up to a point in time; (2) estimating the propensity score of individuals in the risk sets, i.e., the probability of being treated given pre-treatment characteristics; and (3) using the propensity score to match individuals who went on to receive the treatment (here, fringe-interactions) with those who did not. We describe this process below.

1. *Creating risk sets* – Each treated user $u_{non-fringe}$ is treated at a different treatment time (i.e., receive a fringe-interactions), where the treatment time corresponds to the day a treated user $u_{non-fringe}$ posted on the subreddit $s_{non-fringe}$ the comment $c_{non-fringe}$ that received the fringe-interaction. We subsample the set of possible control users to those controls that posted a comment $c'_{non-fringe}$ on the same subreddit $s_{non-fringe}$ on the same week of the interaction (treatment). By following this procedure, we create comparable sets of users. We represent each of these users considering the comment-, post-, and profile-related covariates calculated eight weeks before the treatment time t of each risk set.
2. *Training propensity score model* – We compute the propensity to receive an interaction from a fringe user u_{fringe} for each comment $c_{non-fringe}$ written by a treated or control user $u_{non-fringe}$ in a risk set. Specifically, we fine-tune BERT (Devlin et al. 2018) with an additional linear layer to compute the propensity that a comment written by a user receives a fringe-interaction, providing as input to the model the text of the content $c_{non-fringe}$ and the title of the corresponding Reddit submission. Before the final linear layer, we concatenate the [CLS] token with user and post-level covariates (computed considering the user $u_{non-fringe}$ activities in the eight weeks before time t) to compute the propensity of a comment $c_{non-fringe}$ to receive a comment from a fringe user $u_{non-fringe}$.
3. *Matching* – Finally, we match each treated user to a control user (within in each risk set) using the nearest neighbor algorithm on the propensity score considering a calliper of 0.001. Further, for each treated user, we only consider control users who posted in the same subreddit in the same week. This procedure yields 82,725 pairs of users, and there were 3,831 treated users we could not match. We assess the quality of the matching by measuring the standardized mean difference of user-, comment- and post-level covariates, obtaining absolute standardized mean differences smaller than the standard 0.1 threshold (Austin 2011) for most covariates considered (13/15 for r/Incels, 14/15 for r/GenderCritical, and 12/15 for r/The_Donald (see Section 7).

We provide examples of matched comments written by the treated and control users that we match according to our matching procedure in Table 2.

Regression Analysis. Considering matched users, we estimate the effect of fringe-interactions with the linear model

$$y_{ut} = \beta_0 + \beta_1 \text{treated}_{ut} + \boldsymbol{\gamma}^T \mathbf{X}_u + F_t, \quad (2)$$

where y_{ut} represents a binary variable indicating whether a *non-fringe* user u joined a fringe community in the eight weeks after the fringe-interaction at week t; treated_{ut} indicates if user u received the treatment (fringe-interaction) in week t. Therefore, β_1 captures the effect of the fringe-interaction on the non-fringe user u. $\mathbf{X}_u$ represents the array of control variables introduced in Section 3 at the user level computed in the pre-treatment period; and F_w are weekly fixed effects, which control for potential latent time trends that could impact our results. Note that, coefficients here must be interpreted as percentage point (*pp*) increases.

4 Results

RQ1: Effect of Fringe-Interactions

Overall Interaction Effect. Non-fringe users who receive a fringe-interaction exhibit a significant increase in the probability of joining the said fringe community relative to those

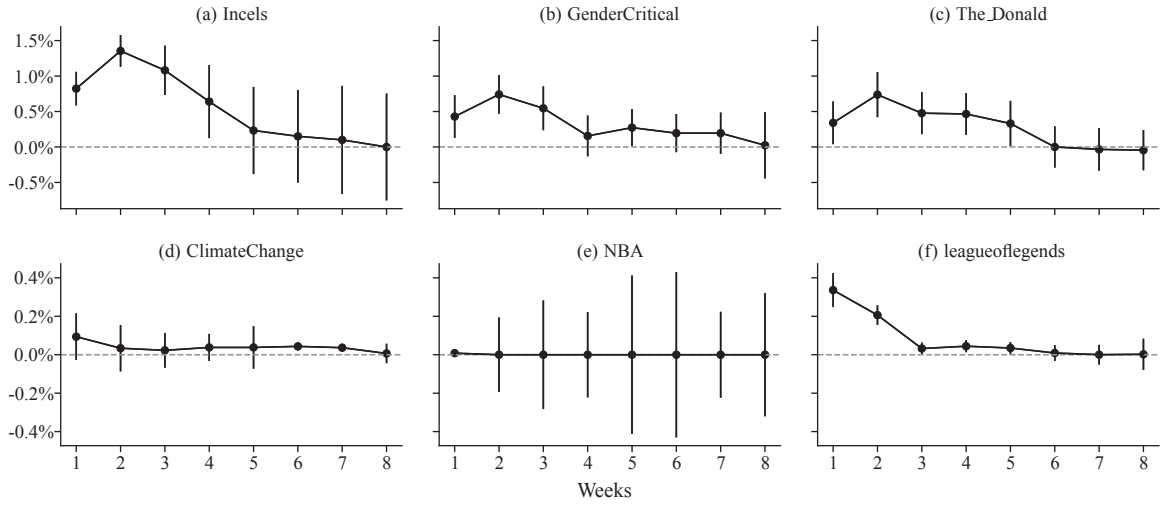

Figure 4: Top row shows the effect size of fringe-interactions on users that interacted with members of fringe communities (r/Incels, r/GenderCritical, r/The_Donald). Bottom row shows the effect of interactions when users interact with members of non-fringe communities (r/climatechange, r/NBA, r/leagueoflegends). The effect is stratified over a follow-up period of eight weeks (x-axis). Note that the y-axis scale is different between top and bottom row.

who do not (see Figure 3). Specifically, considering the regression analysis described in Equation (2), we find that fringe-interactions with r/Incels users increase by 4.2 pp the likelihood of posting on r/Incels in the eight weeks following the interaction with a fringe user of r/Incels (2.4 pp for r/GenderCritical; 2.2 pp for r/The_Donald). To understand if such an effect is specific to fringe communities, we repeat the same analysis considering interactions with users of non-fringe communities. Practically, we substitute the three fringe subreddits with r/climatechange, r/NBA, and r/leagueoflegends. We then repeat our observational study. We find effects to be smaller and not statistically significant, 0.5 pp for r/ClimateChange and r/leagueoflegends; and 0.0006 pp for r/NBA.

Time-Stratified Interaction Effect. To study the variation of the effect within the observation window, we re-run Equation (2) considering the data associated with each week in the post-intervention follow-up period, i.e., we estimate the effect of the user joining the community on the first, second, third week and so on after the fringe-interaction. For the fringe communities considered (Figure 4; top row), we find that the effect is strongest during the second week following the interaction with fringe users (1.3 pp for r/Incels; 0.7 pp for r/GenderCritical; 0.8 pp for r/TheDonald; all results statistically significant with $p < 0.05$). This likely happens because the very first week can encompass fewer days in our setup (e.g., if the external interaction happened on a Friday, fewer days are being considered in week #1). After the second week, the effect gradually wanes for all three communities. For the non-fringe communities considered (Figure 4; bottom row), weekly effects are only significant for the r/leagueoflegend community, which, interestingly, does not exhibit the peak in the second week.

RQ2: In Which Subreddits Are Fringe-Interactions Successful?

To investigate which characteristics make subreddits more susceptible to the effect of fringe-interactions , we categorize subreddits across three "social dimension scores" as computed by Waller and Anderson (2021). These represent the social positioning based on the partisanship, age, and gender of a subreddit. Each score is a value between -1 and $+1$, describing how left or right, old or young, and how feminine or masculine a subreddit is.

We assign subreddits where a fringe-interaction occurred in one of ten groups (D1-D10). These groups are obtained by (1) computing the ten deciles of the social scores and (2) assigning each subreddit to one of these groups based on its social score. We repeat this procedure for all three social dimensions. To study in which subreddits fringe-interactions are most effective, we run the following regression:

$$y_u = \beta_0 + \beta_1 \text{treated}_u + \boldsymbol{\beta}^T \boldsymbol{D}_{s(u)} \text{treated}_u + \boldsymbol{\gamma}^T \mathbf{X}_u + F_t \quad (3)$$

here $D_{s(u)}$ is a one hot encoded vector that is zero everywhere but in the position of the decile assigned to the subreddit s where the non-fringe user u received the fringe-interaction. We index this subreddit with $s(u)$. Therefore, the vector of coefficients $\boldsymbol{\beta}$ captures the effect of an interaction occurred in a subreddit associated in a specific group (D1-D10). The other variables are as described in Equation (2).

Considering the partisan dimension, interactions with r/Incels and r/The_Donald users are most successful in right-leaning subreddits (15.3 pp D9 for r/Incels) and (2.5 pp D9 for r/The_Donald interactions). For example, interactions are effective in subreddits like r/GunsForSale, r/russia, and

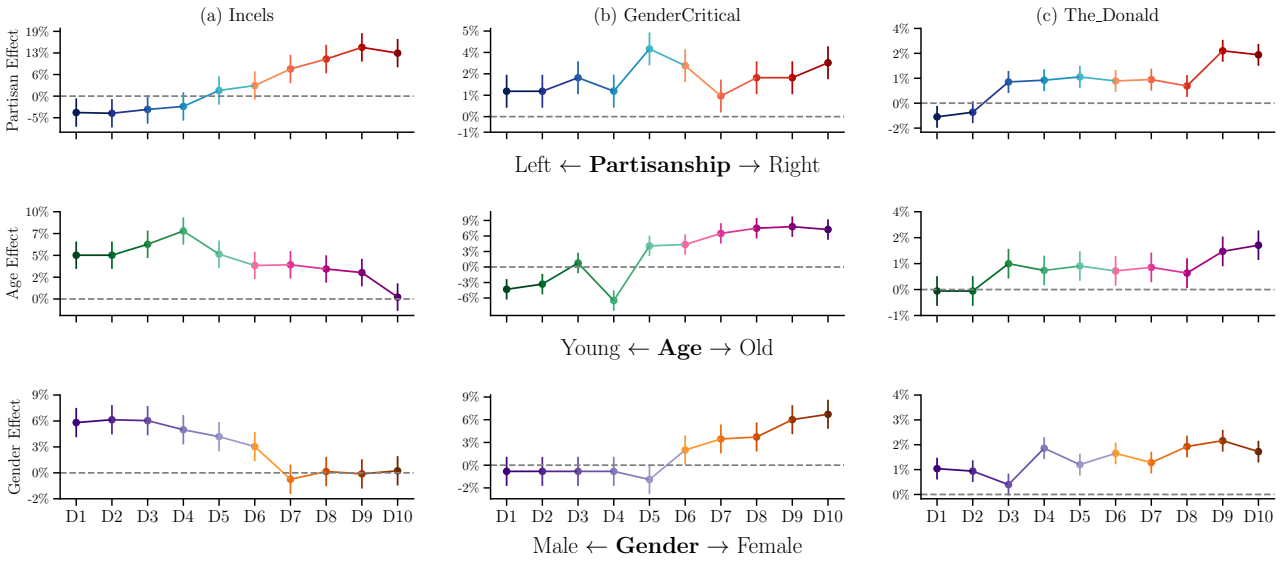

Figure 5: *Stratification of External Interaction Effects Based on Social Dimension Scores.* Effects of fringe-interactions are categorized according to the partisan (top row), age (middle row), and gender (bottom row) scores of the subreddits where the fringe-interaction occurred. Subreddits are grouped into ten distinct categories (D1-D10), see Table 3, according to their social dimension scores (x-axis). Each column illustrates the effect of fringe-interactions with users of r/Incels, r/GenderCritical, or r/The_Donald. All results statistically significant with $p < 0.05$

r/totalwar which are part of D9 group. Differently, fringe-interactions with r/Incels and r/The_Donald users are least successful in left-leaning subreddits, where we found statistically significant negative effects. Examples of subreddits in the D1 and D2 along the partisan scores are r/democrats, r/Marijuana, and r/Impeach_Trump. No similar political divide is observed for r/GenderCritical.

For the age dimension, interactions with r/Incels users are most successful on "young" subreddits (D1-D4) with an average effect of 6.2 pp, whereas for interactions with r/GenderCritical and r/The_Donald users, the effect is more pronounced in "older" subreddits (D5-D10) with average scores of 6.6 pp and 1.5 pp, respectively. Interestingly, the effect on young subreddits is negative for r/GenderCritical (-3.3 pp) and negligible for r/The_Donald (0.06 pp) .

For the gender dimension, interactions with r/Incels users show higher effects in masculine subreddits, while r/GenderCritical interactions have a higher impact on feminine subreddits. No clear trend is observed for interactions with r/The_Donald users.

RQ3: The Effect of Linguistic Traits

Last, we investigate whether linguistic traits of fringe-interactions (i.e., comments from fringe to non-fringe users) impact their effectiveness in attracting newcomers with the following linear model:

$$y_u = \beta_0 + \beta_1 \text{treated}_u + \boldsymbol{\beta}^T \boldsymbol{L_{c(u)}} \text{treated}_u + \boldsymbol{\gamma}^T \mathbf{X}_u + F_t \quad (4)$$

where $L_{c(u)}$ represents the vector of linguistic traits of the comment c written by the fringe user and received by the non-fringe user u in the fringe-interaction (we index this comment with $c(u)$). The vector of coefficients $\boldsymbol{\beta}$ represents the effect of each linguistic trait considered. The other variables are as described in Equation (2) and Equation (3).

Specifically, the linguistic traits we consider are (1) anxiety, sadness, and "they vs. we" language computed accordingly to LIWC (Pennebaker, Francis, and Booth 2001), (2) toxicity as computed by Perspective's API (Jigsaw 2022) and binarized by considering texts with a score above 0.8 as toxic (similarly to Ribeiro et al. (2021)), and (3) community-specific lingo (see Section 7). We chose these linguistic traits because involvement in fringe groups is linked to an underlying psychological disposition that includes tendencies to experience emotions such as sadness and anxiety, as well as receptiveness to fringe narratives (e.g., community lingo and highly toxic content) (Chandrasekharan et al. 2017; Butter and Knight 2020). We code all these traits to assign values of 1 if the linguistic trait is present in the interaction and 0 otherwise.

Figure 6 shows the observed effects for selected linguistic traits. Toxic fringe-interactions increase the attraction of newcomers by 5.8, 4.2, 6.1 pp for r/Incels, r/GenderCritical, and r/The_Donald, respectively. Also, we find that comments containing anxious messages and "they vs. we" have a positive effect on attracting newcomers. Interestingly, for r/Incels, we find a negative effect for fringe-interactions containing community-specific lingo.

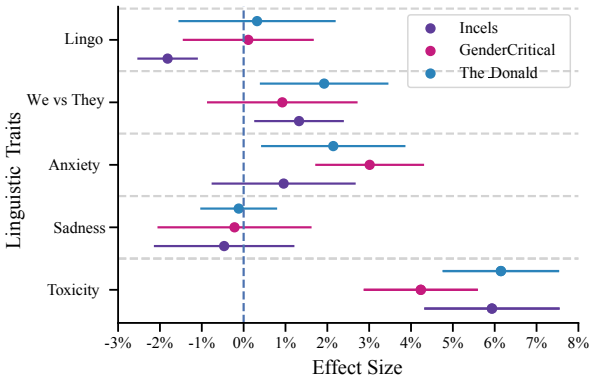

Figure 6: The effectiveness of linguistic traits of fringe-interactions — On the y-axis, the linguistic traits considered. On the x-axis, the effect size of each linguistic traits associated with each fringe subreddit.

5 Robustness Checks

Content-Level Exact Matching

Our findings rely on the semantic similarity of comments we matched using the approach described in section 3. To ensure the validity of our results, we established a strict criterion that the content of treated and control interactions (i.e., comments) must be exactly the same. To fulfil this requirement, we collected 2,312 treated (received a fringe-interaction) and 5,627 control comments with the same hyperlink as their sole content. Due to possible user and post-level confounders, we matched comments that received an interaction with those that did not, taking into account all user and post-level covariates. We obtain $2,274$ pairs of matched comments considering fringe-interactions from r/Incels, r/GenderCritical, and r/The_Donald together. We run our regression analysis, finding average treatment effect of 1.3 pp with ($p < 0.001$).

Sensitivity Analysis

Our results rely on the assumption that the treatment assignment is not biased. Meaning that the only difference between treated and control users is a simple coin flip. Sensitivity analysis is a way of quantifying how the results of our study would change if this assumption is violated (Rosenbaum 2005), This notion is quantified by the sensitivity Γ, which specifies the ratio by which the treatment of two matched persons would need to differ to result in a p-value above the significance threshold. Large values of Γ corresponding to more robust conclusions. For the chosen $p = 0.05$, we measured for the three fringe communities that we study Γs of 2.3, 1.35, and 1.65, which implies that, within matched pairs, an individual's probability of being the treated one could take on any value between $1/(1+\Gamma) = 0.3$ and $\Gamma/(1+\Gamma) = 0.69$ for r/Incels, between 0.42 and 0.57 for r/GenderCritical, and between 0.38 and 0.62 for r/The_Donald without changing our decision of rejecting the null hypothesis of no effect.

6 Discussion

Understanding how fringe communities grow holds immense significance for mainstream platforms and policymakers alike – it is a stepping stone for interventions aimed at limiting their (well-documented) harm. While previous work has pointed at algorithmic visibility and platform affordances as means by which fringe communities grow (**?**), here we hypothesize and study another, more social mechanism. We show that users who involuntarily receive interactions from fringe users increase their likelihood of participating in their fringe communities, and that this effect is modulated by where the interaction happens and what is said by the fringe user.

Implications. Our results raise an important question: Is this mechanism relevant enough to warrant the attention of moderation policies? Within the observation period considered, $40,321$, $32,306$, and $123,562$ users joined r/Incels, r/GenderCritical, r/The_Donald, respectively. Based on our estimated effects, approximately 7.2%, 3.1%, and 2.3% of the newcomers joined after interacting with users from r/Incels, r/GenderCritical, or r/The_Donald. This observation suggests that community-level moderation policies could be combined with sanctions applied to individual users. These sanctions might include reducing the visibility of their posts or limiting the number of comments they can make in more susceptible communities. Such a combination could diminish the impact of fringe-interactions and slow down the growth of fringe communities on mainstream platforms.

Limitations and Future Work. As our conclusions rely on observational data, potential confounders could limit the validity of our study. However, the robustness checks we conducted (content exact matching, sensitivity analysis) mitigate potential threats to our study's validity. Moreover, while our analysis centers on Reddit, where the subreddit structure offers an ideal context for our study, it's worth noting that prior research has stressed the significance of comprehending these mechanisms across various platforms (Horta Ribeiro et al. 2021). Future studies could extend this analysis to platforms lacking distinct communities, potentially yielding broader insight into the spread of fringe ideologies online.

Broader Impact, While our results suggest that curtailing fringe-interactions may effectively reduce the growth of fringe communities on mainstream platforms, it raises ethical concerns about potential censorship. We make two observations on that matter. First, the ethical implications of limiting specific platform interactions deserve careful consideration. Balancing the need to mitigate harmful fringe communities with respect for free expression is a complex challenge, and platform stakeholders must weigh the potential harm of fringe communities against the principles of free speech. Second, the narrative of "free speech" is a crucial mechanism that fringe platforms, such as Parler, use to lure users away from mainstream platforms. Limiting users' ability to post may result in migrations toward fringe platforms, putting users at risk of further radicalization.

References

Ammari, T.; and Schoenebeck, S. 2015. Understanding and supporting fathers and fatherhood on social media sites. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 1905–1914.

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3): 399–424.

Backstrom, L.; Kumar, R.; Marlow, C.; Novak, J.; and Tomkins, A. 2008. Preferential behavior in online groups. In *Proceedings of the 2008 international conference on web search and data mining*, 117–128.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The PushShift Reddit dataset. In *'20Proceedings of the international AAAI conference on web and social media*, volume 14, 830–839.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Butter, M.; and Knight, P. 2020. *Routledge handbook of conspiracy theories*. Routledge.

Caren, N.; Jowers, K.; and Gaby, S. 2012. A social movement online community: Stormfront and the white nationalist movement. In *Media, movements, and political change*, 163–193. Emerald Group Publishing Limited.

Chandrasekharan, E.; Jhaver, S.; Bruckman, A.; and Gilbert, E. 2022. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4): 1–26.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–22.

Collins, B.; and Zadrozny, B. 2020. Facebook bans QAnon across its platforms. https://www.nbcnews.com/tech/tech-news/facebook-bans-qanon-across-its-platforms-n1242339.

Corso, F.; Russo, G.; and Pierri, F. 2024. A Longitudinal Study of Italian and French Reddit Conversations Around the Russian Invasion of Ukraine. *arXiv preprint arXiv:2402.04999*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dodge, J.; Ilharco, G.; Schwartz, R.; Farhadi, A.; Hajishirzi, H.; and Smith, N. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

Gelman, A. 2014. It's not matching or regression; it's matching and regression.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Hoffman, B.; Ware, J.; and Shapiro, E. 2020. Assessing the threat of incel violence. *Studies in Conflict & Terrorism*, 43(7): 565–587.

Horta Ribeiro, M.; Jhaver, S.; Zannettou, S.; Blackburn, J.; Stringhini, G.; De Cristofaro, E.; and West, R. 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): 1–24.

Hwang, S.; and Foote, J. D. 2021. Why do people participate in small online communities? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–25.

Jaki, S.; De Smedt, T.; Gwóźdź, M.; Panchal, R.; Rossa, A.; and De Pauw, G. 2019. Online hatred of women in the Incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2): 240–268.

Jhaver, S.; Boylston, C.; Yang, D.; and Bruckman, A. 2021. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the CSW2'21 on Human-Computer Interaction*, 5(CSCW2): 1–30.

Jigsaw. 2022. Perspective API. https://perspectiveapi.com/.

Kaitlyn, T. 2020. The Secret Internet of TERFs. https://www.theatlantic.com/technology/archive/2020/12/reddit-ovarit-the-donald/617320/. Accessed on 2022-08-26.

Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; and Tamura, K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, 35(6): 1547.

Lampe, C.; and Johnston, E. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 ACM International Conference on Supporting Group Work*, 11–20.

Lauckner, C.; and Hsieh, G. 2013. The Presentation of Health-Related Search Results and Its Impact on Negative Emotional Outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 333–342. New York, NY, USA: Association for Computing Machinery. ISBN 9781450318990.

Lingel, J.; Naaman, M.; and Boyd, D. M. 2014. City, self, network: transnational migrants and online identity work. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 1502–1510.

Lu, X.; Phang, C. W.; and Yu, J. 2011. Encouraging participation in virtual communities through usability and sociability development: An empirical investigation. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 42(3): 96–114.

Lyons, M. N. 2017. Ctrl-alt-delete: The origins and ideology of the alternative right. *Political Research Associates*, 20.

Monti, C.; Cinelli, M.; Valensise, C.; Quattrociocchi, W.; and Starnini, M. 2023. Online conspiracy communities are more resilient to deplatforming. *arXiv preprint arXiv:2303.12115*.

Paudel, P.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; and Stringhini, G. 2021. Soros, child sacrifices, and 5G: understanding the spread of conspiracy theories on web communities. *arXiv:2111.02187*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*.

Phadke, S.; Samory, M.; and Mitra, T. 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*, volume 16, 770–781.

Ren, Y.; Harper, F. M.; Drenner, S.; Terveen, L.; Kiesler, S.; Riedl, J.; and Kraut, R. E. 2012. Building member attachment in online communities: Applying theories of group identity and interpersonal bonds. *MIS quarterly*, 841–864.

Ribeiro, M. H.; Blackburn, J.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; Long, S.; Greenberg, S.; and Zannettou, S. 2021. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 196–207.

Ridings, C. M.; and Gefen, D. 2004. Virtual community attraction: Why people hang out online. *Journal of Computer-mediated communication*, 10(1): JCMC10110.

Rosenbaum, P. R. 2005. Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Russo, G.; Hollenstein, N.; Musat, C.; and Zhang, C. 2020. Control, generate, augment: A scalable framework for multi-attribute text generation. *arXiv preprint arXiv:2004.14983*.

Russo, G.; Horta Ribeiro, M.; Casiraghi, G.; and Verginer, L. 2023a. Understanding online migration decisions following the banning of radical communities. In *Proceedings of the 15th ACM Web Science Conference 2023*, 251–259.

Russo, G.; Stoehr, N.; and Ribeiro, M. H. 2023. Acti at evalita 2023: Overview of the conspiracy theory identification task. *arXiv preprint arXiv:2307.06954*.

Russo, G.; Verginer, L.; Ribeiro, M. H.; and Casiraghi, G. 2023b. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 742–753.

Šakota, M.; Peyrard, M.; and West, R. 2024. Fly-swat or cannon? cost-effective language model choice via meta-modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 606–615.

Samory, M.; and Mitra, T. 2018. Conspiracies online: User discussions in a conspiracy community following dramatic events. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Schmid, A. P. 2013. Radicalisation, de-radicalisation, counter-radicalisation: A conceptual discussion and literature review. *ICCT research paper*, 97(1): 22.

Schmitz, M.; Muric, G.; and Burghardt, K. 2022. Quantifying how hateful communities radicalize online users. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 139–146. IEEE.

Trujillo, A.; and Cresci, S. 2022. Make reddit great again: assessing community effects of moderation interventions on r/the_donald. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2): 1–28.

Waller, I.; and Anderson, A. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, 1954–1964.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature*.

Washington Post. 2020. 'Reddit closes long-running forum supporting President Trump after years of policy violations'.

Weld, G.; West, P.; Glenski, M.; Arbour, D.; Rossi, R. A.; and Althoff, T. 2022. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In *ICWSM*.

Williams, C. 2020. The ontological woman: A history of deauthentication, dehumanization, and violence. *The Sociological Review*, 68(4): 718–734.

Zhang, J.; Hamilton, W.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 377–386.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv:1904.09675*.

Paper Checklist

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? No

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes

(e) Did you describe the limitations of your work? Yes

(f) Did you discuss any potential negative societal impacts of your work? Yes

(g) Did you discuss any potential misuse of your work? Yes

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

1. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Yes

(b) Have you provided justifications for all theoretical results? Yes

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes

(e) Did you address potential biases or limitations in your theoretical framework? Yes

(f) Have you related your theoretical results to the existing literature in social science? Yes

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes

7 Appendix

Finetuning of BERT. We fine-tune BERT (Devlin et al. 2018) with an additional linear layer to compute the propensity that a comment written by a user receives a fringe-interaction. To do so, this model takes as input the content of the comment $c_{non-fringe}$ and the title of

the corresponding Reddit submission. We, then, concatenate the [CLS] token with user and post-level covariates to compute the propensity of a comment $c_{non-fringe}$ to receive a comment from a fringe user $u_{non-fringe}$

We fine-tune this model by following the recommendation of Dodge et al. (2020). Specifically, we balance our training data including all comments $c_{non-fringe}$ written by a non-fringe user $u_{non-fringe}$ that received a fringe-interaction (treatment group) and a subsample of possible controls such as to build a balanced dataset for the training of BERT and the additional linear layer. We then repeat the fine-tuning using five different random seeds for fifty different subsamples of the control group. We do not observe a statistical difference between the values of the loss function for the different subsamples. We fine-tune an instance of this model for each of groups of interactions wit users of r/Incels, r/GenderCritical, and r/The_Donald. We trained each of these instances for 5 epochs in total. A Tesla T4 GPU for the fine-tuning of the model. Finally, Table 3 shows a list of examples of comments written by treated and control users that that we matched using the model described above. Recent advances in NLP can provide future avenues of research to improve the training of models for propensity score matching (Šakota, Peyrard, and West 2024; Brown et al. 2020; Russo et al. 2020)

Robustness of Balanced Risk Set Matching We have evaluated the robustness of our results against two different matching procedure. The first is described in Section 3 which is based on propensity score matching and nearest neighbour algorithm. The second approach we match treatment and control unitis directly in the covariates space. To do so, we compute the cosine similarity between the vectoral representations (i.e., [CLS]-token concatenated with post and user-level covariates) and match them using the nearest neighbor algorithm. We do not find any statistically significant difference in our results using these two approaches. To show the quality of our matching, we show in Figure 7 "love plots" for the absolute standardized mean differences of all post and user profile covariates described in Section 3. To further check the average similarity of our matched comments at semantic level considering we measure the average BERTScore (Zhang et al. 2019) of the treatment comments with (i) the matched control comments and (ii) ten random samples of all possible control comments. Moreover, we also consider other linguistic traits like toxicity and LIWC dimensions to obtain a more interpretable representation of the similarity between matched comments. We show the values of the absolute standardized mean errors (ASMD) before and after matching for all three subreddits we analyzed.

Subreddits Social Dimensions In Section 4, we stratify the effect of fringe-interactions across three "social dimensions" representing subreddits partisanship, age and gender. We assign each subreddit to one of ten groups based on its social dimension score (deciles). Table 3 (bottom) provides a list of examples of subreddits in each of the decile for partisan, age and gender score computed accordingly to Waller and Anderson (2019).

Collection of Communities Lingo. To collect the lingo of r/Incels, r/GenderCritical, and r/The_Donald specific community lingo we build a specialized web-crawler. We then collected from the offical websites of these three communities the list of the terms creating the glossary of these three communities. We then count for each of the comments in our analysis the number of Incels, GenderCritical, and The_Donald related terms.

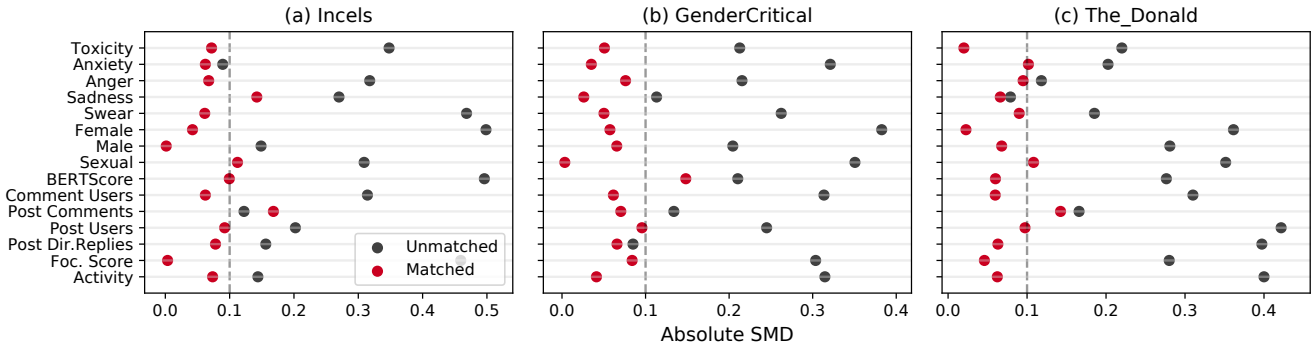

Figure 7: Absolute standardized mean differences (SMD) for fringe interactions before and after matching for all the covariates used for the balanced risk set matching. The SMDs have been computed for all three considered subreddits r/Incels, r/GenderCritical, and r/The Donald

Treatment	Control
Maybe you are dressing or groomed so you look somewhat gay and why women aren't going out with you and men are hitting on you. Just a thought, no disrespect intended.	The whole point of being gay is that you are attracted to other men. Not men dressed as women or people born a woman and identify as a man and use a strap on. Sorry honey but maybe try Tinder
China's top politicians are mainly scientists, The USA's top politicians are mainly religious zealots and dumbfucks. Guess who invests more in climate change.	If anyone is going to speak of climate change, they must be armed with the true facts. Unfortunately, the major data points are very skewed. This was uncovered via Wikileaks. Follow the money.
Trump is the symptom of everything that is wrong with US politics while Hillary is the cause. Real talk that was worded beautifully, I wholeheartedly agree.	You're welcome, but please enlighten me on why Hillary should be president without mentioning Donald Trump.
If survival requires that we keep the environment to a certain level of stability, then we must. And what is your stance in the current pollution situation as well as big oil industry?	Yeah, because oil and gas companies are totally struggling for profits - notwithstanding the recent and temporary drop in oil prices. Also, you're aware that a focus on renewables would increase supply, which would decrease prices? It balances out and is certainly better for the environment.

Table 2: Examples of comments that received a fringe-interaction (treated) matched with comments that did not (control)

Decile	Partisan	Age	Gender
D1	democrats, Marijuana	knifeparty, teenagers	malelifestyle, fuckingmanly
D2	drugstore, TwoXSex	skateboarding, Jokes	SRSMen, Machinists
D3	TVDetails, help	uncharted, studyAbroad	DestructionPorn, Archery
D4	see, giantbomb	ImaginaryMonsters, GirlTalk	porngifs, adorableporn
D5	happygirls, MOMs	Xsome, kickasstorrents	AdPorn, xxxcaptions
D6	Magic, porn	SugarBaby, KingstonOntario	weirdal, CryptoKitties
D7	DebateCommunism, ideas	exmormon, MadMax	PalaceClothing, WWE
D8	FTC, BitcoinBeginners	LifeProTips, vermont	gaymers, GirlsXBattle
D9	Gunsforsale, russia	Plumbing, tuckedinkitties	DDLC, BattleCatsCheats
D10	Conservative, Military	RedditForGrownups, MealPrepSunda	againstmensrights, TheGirlSurvivalGuide

Table 3: Examples of subreddits in each decile according to Partisan, Age and Gender Scores. Top rows are most left-leaning (■), young (■), male (■) subreddits, respectively. Bottom rows are most right-leaning (■), old (■), and female (■) subreddits, respectively.