Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Non-polar Opposites: Analyzing the Relationship between Echo Chambers and Hostile Intergroup Interactions on Reddit

Alexandros Efstratiou[1], Jeremy Blackburn[2], Tristan Caulfield[1], Gianluca Stringhini[3], Savvas Zannettou[4], Emiliano De Cristofaro[1]

[1]University College London
[2]Binghamton University
[3]Boston University
[4]Delft University of Technology
alexandros.efstratiou.20@ucl.ac.uk, jblackbu@binghamton.edu, t.caulfield@ucl.ac.uk, gian@bu.edu, s.zannettou@tudelft.nl, e.decristofaro@ucl.ac.uk

Abstract

Previous research has documented the existence of both online echo chambers and hostile intergroup interactions. In this paper, we explore the relationship between these two phenomena by studying the activity of 5.97M Reddit users and 421M comments posted over 13 years. We examine whether users who are more engaged in echo chambers are more hostile when they comment on other communities. We then create a typology of relationships between political communities based on whether their users are toxic to each other, whether echo chamber-like engagement with these communities is associated with polarization, and on the communities' political leanings. We observe both the echo chamber and hostile intergroup interaction phenomena, but neither holds universally across communities. Contrary to popular belief, we find that polarizing and toxic speech is more dominant between communities on the same, rather than opposing, sides of the political spectrum, especially on the left; however, this mostly points to the collective targeting of political outgroups.

1 Introduction

In *echo chambers*, users encounter view-affirming information or other users, thus never experiencing any informational disruption (Sunstein 2001). Users tend to engage with communities that politically align with their views (Waller and Anderson 2021), while controversial events often lead to spontaneously formed polarized networks (Barberá et al. 2015; Del Vicario et al. 2017). At the same time, users who try to bridge opposing views tend to receive lower attention and social rewards (Garimella et al. 2018). Diversifying user exposure and diminishing such echo chamber spaces is a promising approach to securing the integrity of deliberative democracy (Matakos et al. 2022).

Hostile intergroup interactions pose a challenge to this approach. Participation in mixed social media networks (Vaccari et al. 2016), diverse media diets (Guess et al. 2018), and encountering disagreeable views (Dubois and Blank 2018) are all fairly common. When interactions between users on opposing camps do occur, however, they tend to be more toxic and hostile (De Francisci Morales, Monti, and Starnini 2021; Cinelli et al. 2021; Bliuc, Smith, and Moynihan 2020; Marchal 2022).

Problem statement. We hypothesize that echo chambers and hostile interactions may not be mutually exclusive. Instead, one may influence the degree to which the other occurs. Nonetheless, this has yet to be explored.

Moreover, users may display varying degrees of engagement with their "echo chambers." Research that analyzes echo chambers at the community level may thus not capture this. Here, we set out to recognize such understudied differences in engagement using a user-level approach.

Overall, we focus on two main research questions:

RQ1 How is a user's degree of engagement with a political echo chamber related to their hostility in an intergroup interaction?

RQ2 What are the different relationships between Reddit political communities based on the hostility and the polarization of their user base, and how do they vary depending on their political leanings?

Methodology. Our work builds on a dataset of 421M comments made between 2006 and 2019 from 5.97M unique authors on Reddit. These appeared across 918 political subreddits, which we cluster into distinct "echo chamber" communities based on their user similarities (see Section 3). We allocate users into home communities by analyzing where they were most active over these 13 years and measure the toxicity of the comments left by each community's home users on each of the other communities (Section 4).

For this study, any community can be an echo chamber for a given user if they only (or disproportionately) engage with it. Therefore, we define echo chamber engagement as the proportion of comments that a user left in their preferred community, reflecting their preference for homophily (i.e., their tendency to surround themselves with similar others (Rogers and Bhowmik 1970)). This deviates somewhat from traditional definitions of echo chambers (Sunstein 2001) as it is content-agnostic, but it is in line with other work attempting to quantify echo chambers at scale (Brugnoli et al. 2019; Waller and Anderson 2021; Zollo et al. 2017). We define an intergroup interaction as the event of a user commenting on a community other than their echo

chamber, where the interaction is hostile if the comment is toxic (as determined through Google's Perspective API).

To address **RQ1**, we use mixed-effects logistic models to assess how the probability that a user's comment will be toxic on some target community is related to the proportion of comments left by the user in their echo chamber (we do so for every possible community pair). We then combine our cross-toxicity and mixed-effects analyses to create a typology of community relationships (Section 5) and observe the frequency of each type based on a manual assessment of the communities' political leanings (i.e., whether they are on the same or opposing sides) to address **RQ2**.

Main findings. Overall, Reddit's political space between 2006-2019 included 16 communities not captured by a binary left-right split. Some of these communities were almost universally toxic (or non-toxic), while others showed more selectivity in *where* they were toxic. We also show that increased engagement with these "echo chamber" communities had differential relationships to hostility outside of them, depending on the target community. Toxic behavior was up to 2.5 times more likely with higher echo chamber engagement when the relationship between communities was polarizing, and down to nearly 70 times less likely when the relationship was depolarizing; however, this depolarizing effect may also be attributable to content moderation.

We do not find universal tribalism on Reddit. Specifically, the most common type of relationship (˜21%) was an indifferent one. Contrary to conventional wisdom (De Francisci Morales, Monti, and Starnini 2021; Marchal 2022), inciting and polarizing relationships were more common between communities on the same (˜6%) rather than on opposing sides (˜2%) of the political spectrum.

Our contributions are three-fold. First, we make a first step toward *systematically* typologizing community relationships. This provides a more accurate map of the state of political discourse, including understudied elements such as indifferent communities, polarizing relationships between communities of similar leanings, and civil relationships between communities of opposite leanings. Second, we reveal that whereas increased engagement with some communities is indeed associated with increased hostility toward others, the opposite relationship holds for several communities. This can allow future research to better target anti-polarization interventions like diversified exposure. For example, increasing network diversity should be a promising approach in cases where higher echo chamber engagement is related to higher hostility but may fail in cases where the opposite is true. Finally, we open future research directions for the role of moderation in these observations.

2 Background and Related Work

In this section, we cover prior work on echo chambers and hostile intergroup interactions online before discussing the research gap in attempting to link these two phenomena.

2.1 Echo Chambers

Echo chambers are fairly widespread on social media (Terren and Borge-Bravo 2021). In terms of the content that users are exposed to, roughly 90% of the political videos that the average user consumes on YouTube align with their political beliefs (Hosseinmardi et al. 2020). Furthermore, science-advocating Facebook users tend to only interact with scientific pages, whereas conspiratorial users only interact with conspiratorial pages; users interacting with both kinds of pages are very rare (Brugnoli et al. 2019). While several fact-checks are aimed toward these conspiratorial users, Zollo et al. (2017) find that only about 1.2% of them interact with this information.

Echo chambers may also arise via interactions with similar users. Garimella et al. (2018) find high polarity in political networks on Twitter, with highly partisan users receiving more engagement. On Reddit, users tend to interact with ideologically similar communities (called *subreddits*) (Waller and Anderson 2021). However, De Francisci Morales, Monti, and Starnini (2021), looking at *r/politics* which is one of the largest political subreddits during the 2016 election, find that cross-cutting interactions are fairly common there.

Specific platform affordances may play a role in the formation of echo chambers. In a platform comparison study, Cinelli et al. (2021) find that echo chambers are more prominent on Facebook and Twitter than on Reddit. This may be because Facebook and Twitter make heavier use of recommender algorithms, resulting in so-called *filter bubbles* (Pariser 2011). Indeed, Bakshy, Messing, and Adamic (2015) find that introducing algorithmic ranking of content can reduce the exposure of Facebook users to cross-cutting content, although individual user choice has a larger effect on this. On Spotify, recommendations can reduce the overall diversity of podcasts that individual users engage with (Holtz et al. 2020). A simulation study finds that several different types of recommender algorithms can increase the similarity of content that already similar users engage with (Chaney, Stewart, and Engelhardt 2018).

Echo chambers may be exacerbated by controversy around a given topic. Controversial events which cause nationwide debates eventually lead to echo chamber discussions between users of similar beliefs (Barberá et al. 2015; Del Vicario et al. 2017). Radicalization through similar content exposure is another factor; for example, Hosseinmardi et al. (2020) report a surge in alt-right video consumption on YouTube, with radicalization occurring only for right-wing users. Similarly, Ribeiro et al. (2020) show that initial consumption of mild right-leaning content can lead to eventual consumption of far-right content.

Overall, echo chambers may alienate users to certain points of view, making them apprehensive of such views when they do encounter them. The small fraction of conspiratorial users who interact with fact-checks in Zollo et al. (2017) become more polarized following this exposure. Relatedly, the higher users' activity in their preferred spaces, the more polarized these users become (Brugnoli et al. 2019). Therefore, disproportional interaction with only specific kinds of content or users may affect behavior upon interaction with other kinds.

2.2 Intergroup Interactions

When partisans witness criticism against their side, they wish to distance themselves from opposing partisans (Suhay, Bello-Pardo, and Maurer 2018). However, they also overestimate how much the latter is prejudiced against them (Moore-Berg et al. 2020), and correcting these perceptions can reduce political intergroup prejudice (Lees and Cikara 2020). Thus, engaging with oppositional users can both increase polarization (if the user witnesses criticism) or decrease it (if perceptions of prejudice are corrected); research so far mostly supports the former.

Two separate studies on the r/politics community on Reddit find that cross-partisan interactions are fairly common but tend to be more hostile (De Francisci Morales, Monti, and Starnini 2021; Marchal 2022). A YouTube case study of a controversial video finds that users in the comment section often engage in hostile interactions with users of opposing views (Bliuc, Smith, and Moynihan 2020).

Such interactions are not always naturally occurring. Some Reddit users have "anti-social homes" where they go to display elevated hostility (Datta and Adar 2019). Moreover, Kumar et al. (2018) find that users on certain subreddits initiate negative mobilizations on others by posting links targeting posts in other communities. "Brigading attacks", i.e., targeting another community to down-vote posts and harass its users, also occur on Reddit (Mills 2018).

Hostile intergroup interactions may be elevated during election periods (Datta and Adar 2019), and toxicity is higher when political discussions occur in explicitly political rather than non-political Reddit spaces (Rajadesingan, Budak, and Resnick 2021). The norms of a given community also play a role in the prominence of toxic content there (Rajadesingan, Resnick, and Budak 2020).

Interacting with outgroups can have various effects. Twitter users who are asked to follow bots posting oppositional content become even more entrenched in their prior views 1.5 months later (Bail et al. 2018). Similarly, fact-checks aimed toward conspiratorial Facebook users seem to backfire and drive more conspiratorial content engagement (Zollo et al. 2017). On Reddit, negative interactions with outgroup members reduce the likelihood that such cross-cutting interactions will reoccur in the future (Marchal 2022). On the other hand, sports fans who engage in cross-cutting interactions use more problematic language in their teams' communities (Zhang, Tan, and Lv 2019).

2.3 Remarks

Although several studies have outlined how increased echo chamber engagement may drive higher polarization (Brugnoli et al. 2019) and radicalization (Hosseinmardi et al. 2020; Ribeiro et al. 2020), as well as how cross-cutting exposure may drive higher preference for echo chamber-like consumption (Bail et al. 2018; Barberá et al. 2015; Del Vicario et al. 2017), it remains unclear whether the degree of echo chamber engagement is related to the subsequent hostility expressed in intergroup interactions. To the best of our knowledge, we are the first to study this. We examine the bulk of Reddit's political sphere to understand the dynamics between engagement, hostility, and the political leanings of different communities.

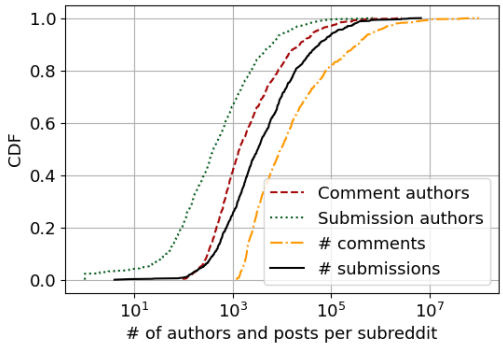

Figure 1: Cumulative Distribution Function (CDF) of the numbers of unique authors and posts for submissions and comments per subreddit.

3 Dataset and Political Communities

In this section, we present our dataset and how we cluster subreddits to identify distinct political communities.

3.1 Data Sources

We start with a list of 31K subreddits, labeled based on the percentage of political comments they host by Rajadesingan, Budak, and Resnick (2021). We treat political subreddits as those hosting 50% or more political content and retain those with at least 1,000 comments and made by at least 100 unique authors. This leaves 918 subreddits.

We obtain all comments posted in these 918 subreddits between June 12th, 2006, and December 31st, 2019, using the Pushshift Reddit dataset (Baumgartner et al. 2020). Due to the last comment in our dataset being made in 2019, our analyses are historical and may not reflect the current state of Reddit. Overall, we examine 421M comments from 5.97M authors. In Figure 1, we plot the Cumulative Distribution Function (CDF) of the number of comments, comment authors, submissions, and submission authors. The normal distributions suggest that the chosen subreddits provide an adequate approximation of Reddit's political sphere across spaces with varying degrees of engagement.

Ethical considerations. This project received ethical approval from UCL's Research Ethics Committee (Project ID: 19379/001). Note that we do not attempt to identify any users appearing in our dataset beyond the use of unique pseudonyms (usernames) to identify comments made by the same user. We only collect and analyze the minimum amount of data required for our research questions.

3.2 Author Similarity Computation

We cluster individual subreddits into larger communities based on their author similarities. If communities share the same users, they might also host the same kinds of opinions, forming *potential* "echo chambers" for this study.

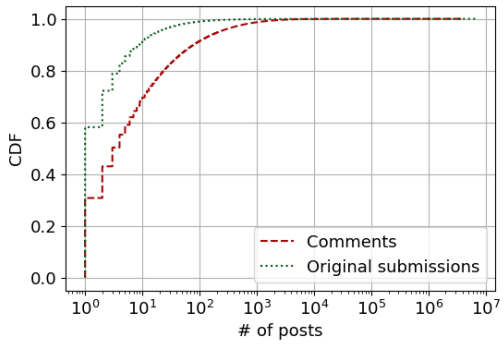

Figure 2: Cumulative Distribution Function (CDF) of comments and submissions across all subreddits per author.

We follow an approach similar to Datta, Phelan, and Adar (2017). First, we create Term Frequency-Inverse Document Frequency (TF-IDF) bag-of-word vectors, where each term is a unique author and each document is an individual subreddit. We then filter out authors with a TF (number of comments) of less than 10 (we choose 10 informed by Figure 2, which shows that approximately 70% of authors post fewer than 10 comments). We also filter document frequency to keep authors who have posted to at least 4 (i.e., median value) and no more than 2.5% (22) of the total subreddits in our dataset. These filters prevent highly sparse vectors and retain only *active* authors whose commenting diversity is informative. The TF-IDF vocabulary, therefore, includes all non-filtered authors. Finally, we compute pairwise cosine similarities between the subreddit TF-IDF vectors.

3.3 Community Detection

Next, as per Datta, Phelan, and Adar (2017), we build a subreddit network using the top 1% cosine similarity values per subreddit as retained undirected edges. We drop the bottom 5% of these edges across *all* subreddits to filter out arbitrary connections (Datta, Phelan, and Adar 2017; von Luxburg 2007) and apply the Louvain algorithm (Blondel et al. 2008) to detect subreddit communities. We obtain a modularity value of 0.58 using this approach, indicating good clustering (Clauset, Newman, and Moore 2004; Newman 2006).

This yields 16 distinct communities; see Table 1. From a manual inspection of the communities, we label 6 communities as *left-leaning* (center-left, pro-Democrat, left-wing, anti-extremism, Socialist, anti-Trump), 5 as *neutral* (EU/UK politics, Middle East/world conflicts, autonews, guns, model politics), and 5 as *right-leaning* (Intellectual Dark Web, pro-Trump, Conservative, alt-right, Libertarian). Figure 3 illustrates the subreddit network retaining nodes with degrees in the top 10%. In Table 2, we provide basic statistics for the overall network and per community. We also provide the full list of the 918 subreddits along with the communities they are allocated to in a Google document.[1]

Note that some of these subreddits have since been banned or restricted, and 24 of them were banned during our observation period. Once again, our analyses reflect a historical, not necessarily a contemporary, picture of Reddit.

To ensure that individual subreddit bans did not have a substantial effect on the aggregated communities, we conduct a time-series analysis where we obtain the Jaccard similarity between the sets of authors who appeared in a given month and its previous month in that community (Figure 4). If subreddit bans drove users out of the entire community, there should be sharp similarity drops following the bans. However, we observe no such drops. Instead, we find somewhat erratic patterns near the start of the communities' formation when the numbers of participating authors were small, followed by convergence toward consistent similarities as the communities grew. The sharp drop in the pro-Trump community around the start of 2016 is also attributable to the growth in authors and activity (which we do not show here due to space constraints). Generally, similarity values for every community, including those with and without banned subreddits alike, remained relatively high (above 0.6). This shows that community participants continued to be active in other subreddits belonging to that community following the bans, and bans did not have substantial effects.

3.4 User Commenting Prevalence

Next, we compute users' commenting prevalence, across the 16 communities, as the proportion of comments they have posted to that community. We note that these prevalence values may be sensitive to cases where a substantial proportion of a user's comments is removed by community moderators. We assume that most users will engage predominantly with the communities they are part of, consistent with findings around homophily on social media (Garimella et al. 2018; Zollo et al. 2017). Thus, we derive each user's "home" community by taking the largest out of the 16 prevalence values (the majority community) for that user.

To filter out "troll" users who post spam or frequent communities with malicious intent (e.g., to harass or provoke), we only consider a user resident if their net upvotes are highest within that community *and* are above 1 (the default score of a newly posted comment). This approach follows An et al. (2019); Rajadesingan, Budak, and Resnick (2021).

4 Toxicity Analysis

This section analyzes how commenting prevalence is related to toxic behavior on Reddit. We focus on how users' involvement in their home community influences how toxic they are elsewhere. Also, we examine whether it is a community's home or non-home users who drive toxic discussions and shed light on toxicity relationships between communities.

Toxicity. We use Perspective API's Severe Toxicity model to label comments as toxic or non-toxic. Severe Toxicity is defined as "a very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective" This provides a score between 0 and 1, and we consider a comment to be toxic if its Severe Toxicity score is above 0.7.

[1]Please see https://docs.google.com/document/d/1XVuHP96zcnrcMqOfEtD3oJ9vmsy8DDG9x8qNYtkz9SU

Community (abbreviation)	Size	Indicative subreddits	#Comments	#Users	#Home users
Center-left (CL)	144	r/politics, r/Liberal, r/obama	236,568,074	4,785,269	2,898,962
Pro-Trump (TR)	68	r/The_Donald, r/Infowars	47,776,673	735,728	256,914
EU/UK Politics (EU-UK)	38	r/unitedkingdom, r/europeans	33,353,351	707,332	243,473
Socialist (SOC)	89	r/MurderedByAOC, r/SandersForPresident	17,429,739	575,307	105,803
Middle East/World conflicts (ME)	71	r/Israel, r/antiwar	15,994,336	683,431	137,894
Anti-Trump (NoTR)	85	r/The_Mueller, r/MarchAgainstTrump	13,802,750	910,826	132,073
Libertarian (LIB)	50	r/Libertarian, r/ronpaul	13,553,498	483,583	76,577
Pro-Democrat (DEM)	34	r/hillaryclinton, r/JoeBiden	9,375,513	184,014	21,390
Left-wing (LEFT)	63	r/communism, r/BlackLivesMatter	7,740,333	445,084	77,293
Anti-political extremes (NoEX)	21	r/stupidpol, r/InternetHitlers	6,584,089	312,920	27,443
Conservative (CON)	27	r/Republican, r/Conservative	5,623,669	251,153	26,619
Intellectual Dark Web (IDW)	63	r/JordanPeterson, r/daverubin	5,563,082	321,540	69,885
Alt-right (ALTR)	48	r/new_right, r/WhiteNationalism	2,835,060	174,035	23,438
Gun discussions (GUN)	25	r/GunsAreCool, r/GunResearch	2,524,063	134,913	24,439
Automatic News (AUTO)	45	r/GUARDIANauto, r/Fox_Nation	1,278,417	43,575	3,773
Model politics (MOD)	40	r/ModelUSGov, r/MHOC	920,556	18,523	4,454

Table 1: List of communities after community detection. Size refers to # subreddits clustered in the respective community.

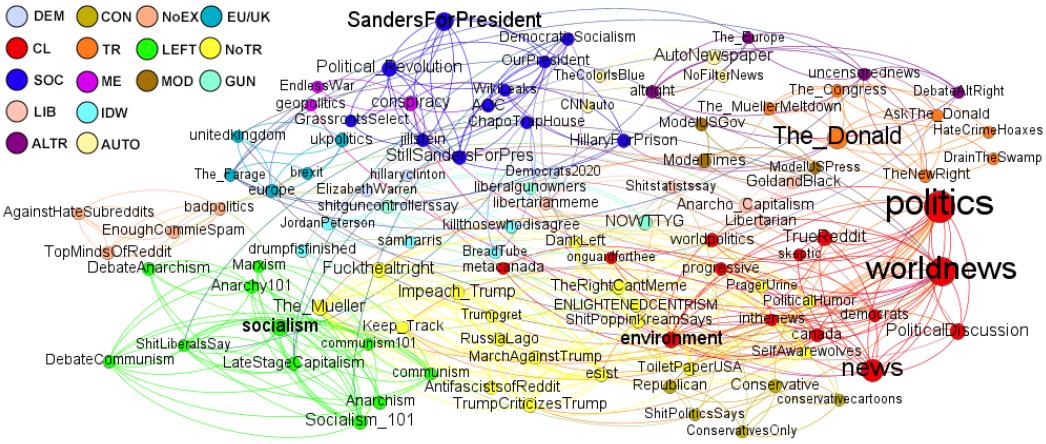

Figure 3: Similar subreddit network with top 10% nodes in terms of degree. Legend shows community that nodes belong to.

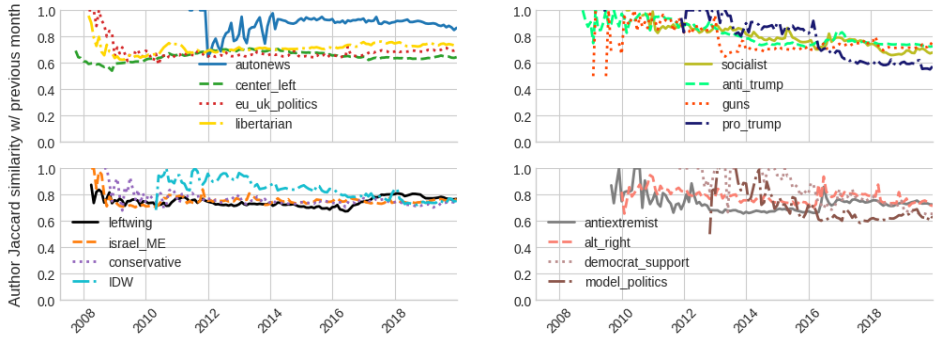

Figure 4: Time-series of Jaccard similarity between participating authors of any given month and the previous month.

Although not free from important limitations, e.g., sensitivity to adversarial text (Jain et al. 2018; Hosseini et al. 2017) and bias toward text mentioning marginalized groups or written in African-American English (Sap et al. 2019), Perspective outperforms alternative models (Zannettou et al. 2020) and allows us to measure relative toxicity at scale.

4.1 Predictors of Community Toxicity

First, we examine the overall toxicity of both home and non-home users in focal communities, as well as the toxicity of

Name	Av. Deg.	D	Dens.	Av. C	Av. PL
CL	8.26	6	0.058	0.35	2.64
TR	6.88	5	0.103	0.42	2.40
EU-UK	8.74	4	0.236	0.60	1.99
SOC	8.63	5	0.098	0.44	2.48
ME	6.17	6	0.088	0.37	2.99
NoTR	8.38	6	0.100	0.41	2.56
LIB	7.00	6	0.143	0.58	2.50
DEM	5.94	5	0.180	0.57	2.49
LEFT	8.76	5	0.141	0.54	2.26
NoEX	9.43	3	0.471	0.79	1.65
CON	6.30	5	0.242	0.65	2.19
IDW	5.97	6	0.096	0.40	2.91
ALTR	6.50	7	0.138	0.49	2.73
GUN	7.44	4	0.310	0.71	2.09
AUTO	7.87	4	0.179	0.54	2.16
MOD	9.80	6	0.251	0.66	2.13
Overall	**12.10**	**6**	**0.013**	**0.25**	**3.29**

Table 2: Network statistics per community and overall. In order, the column names correspond to: Name of the community, average degree, diameter, density, average clustering coefficient, and average path length.

home users in other communities. For every user, we calculate the proportion of toxic comments on each community they have posted. We exclude users with fewer than 10 total comments or fewer than 5 comments on the respective target community to preserve variability. We plot this in Figure 5.

Model politics (MOD) was the least toxic across all three groups of users. Alt-right (ALTR) drew the highest toxicity from both home and non-home users. Anti-Trump (NoTR) users were the most toxic in other communities. In nearly all communities, except center-left (CL), pro-Trump (TR), and socialist (SOC), non-home users were at least as toxic, if not more, as home users.

The green bars in Figure 5, which represent the toxicity of home users in other communities, show that the same users may change their toxic behavior depending on which community they are posting in. For all but three communities (DEM, Intellectual Dark Web (IDW), GUN), these bars are either higher than or lower than *both* home and non-home users' toxicities (i.e., they are not higher than one and lower than the other). This means that in ALTR, anti-extremist (NoEX), and auto-news (AUTO), the most toxic communities overall, users modified their behavior when posting elsewhere. This could be due to better moderation elsewhere, user self-regulation, or, more likely, a combination of both. On the contrary, left-wing (LEFT) and SOC home users became more toxic when posting elsewhere despite these communities being low on toxicity. Our results suggest that community norms influence toxicity levels beyond individual users' tendencies, consistent with the findings of Rajadesingan, Resnick, and Budak (2020).

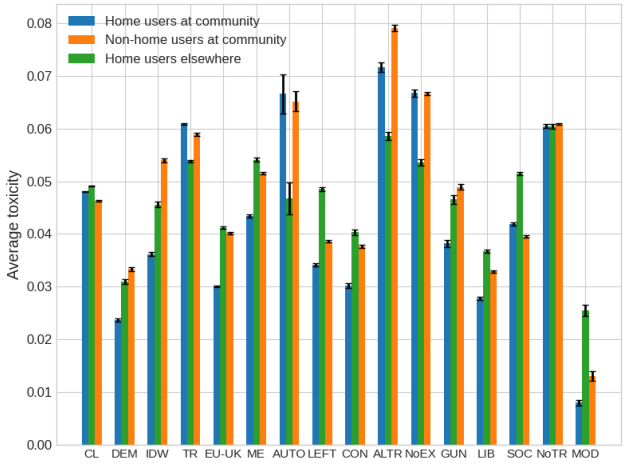

Figure 5: Average toxicity of home and non-home users per community and average toxicity of home users on other communities. Error bars represent standard errors.

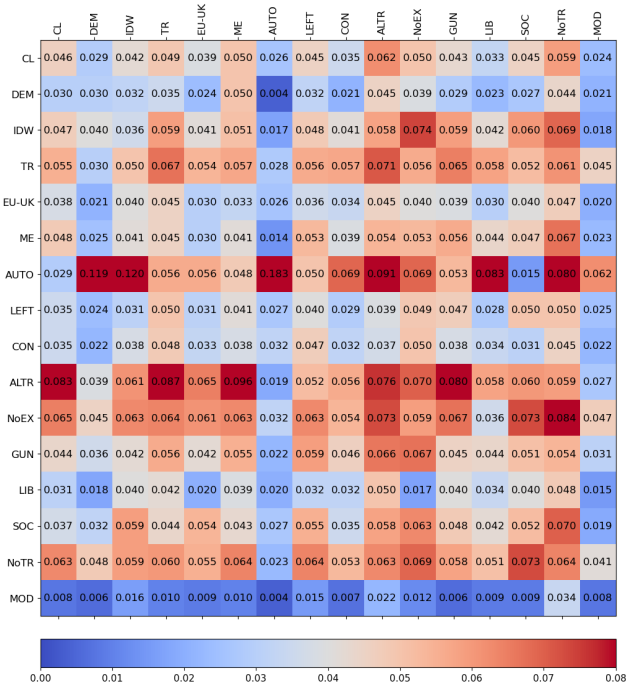

Figure 6: Heatmap showing pairwise proportion of toxic comments posted by the communities' home users. Outgoing toxicity is shown across the horizontal, and incoming toxicity is shown down the vertical.

4.2 Pairwise Community Toxicity

We examine pairwise toxicity relationships between communities. Rather than taking the average toxicity of each home user, we now pool all comments posted from a community's home users on another one and compute the proportion of toxic comments out of the total comments in the pool. Figure 6 is a pairwise toxicity proportion heatmap.

The heatmap follows the user-level toxicity patterns in Figure 5. For example, ALTR and NoEX show much higher toxicity with most comparisons, both in terms of outgoing and incoming toxicity. Similarly, pro-Democrat (DEM) and MOD show lower toxicity across the board.

However, Figure 6 also demonstrates that toxic behavior is not inherently *tribal*. That is, we observe high toxicity between communities on the same side of the political spectrum, e.g., ALTR to pro-Trump (TR) and Socialist (SOC) to NoTR. Similarly, some communities on opposing sides of the political spectrum show lower cross-toxicity—e.g., DEM and Conservative (CON) (in both directions). Nonetheless, moderation potentially plays an important role in these patterns (e.g., selective moderation of toxic comments based on the commenter's political leaning).

4.3 Association of Echo Chamber Engagement with Non-Home Toxicity

Next, we quantify the relationship between echo chamber prevalence and toxicity displayed in other communities. Our goal is to assess, for each possible community pair, whether the posting prevalence of users in the home community was related to the toxicity of their comments at the target.

Modeling. We treat each comment as a single observation. Every comment posted by a user *outside their home community* is a Bernoulli trial, and a successful trial is a *toxic* comment. We then set each user's home community *a posteriori* as described in Section 3.4. Therefore, our model assumes that users do not change their home community over time, which is a potential limitation. We observe each user's comment trail and dynamically update their home posting prevalence based on the number of comments they have posted at home and non-home up to that point.

We treat individual user IDs as nesting variables, maintaining independent observations between users and dependence between comments from the same user. We then run mixed-effects logistic regressions for each pairwise community comparison, allowing the slope and intercept of each user to vary as random effects (Bates et al. 2015):

$$P(toxicity_{c\in T}) = logit(\beta_0 + \beta_1 prevalence_H + b_{0i} + \\ + b_{1i} prevalence_H + \varepsilon)$$

where T is the set of comments in the target community posted from the home community's users in the pairwise comparison, and H is the home community.

Results. We plot all pairwise estimates (β_1 log odds values) in a "faux" forest plot (Figure 7). On average, each comparison consists of 327K comments from 19.5K individual authors. The smallest comparison is 2.01K comments from 308 authors (GUN to DEM), while the largest is 5.97M comments from 390K authors (CL to NoTR). In comparisons where the model fails to converge, we test three different optimization algorithms (nlminb (Gay 1990), L-BFGS-B (Zhu et al. 1997), Nelder-Mead (Nelder and Mead 1965)), and report confidence estimates with successful convergence. Convergence failures are marked with 'cf'.

We omit pairwise comparisons where the home and target community is the same. Furthermore, we omit the AUTO and MOD communities from these analyses because their low numbers of home users do not provide adequate statistical power. We report significance at three different levels: 1) an $\alpha = 0.05$ cutoff, 2) a Bonferroni-corrected $\alpha = 0.05/13$ cutoff for multiple comparisons using the same population (since each population of home users is used 13 times in comparisons), and c) a hyper-conservative $\alpha = 0.05/182$ cutoff point for all comparisons in the plot. We use level a) for our interpretations in the remainder of this paper as we are interested in the unique relationships between each pair, although level b) may also be reasonable. Level c) is only used for transparency purposes, and we do not recommend it as it is over-corrective and can inflate Type II errors.

Discussion. Positive values show "polarizing" associations (i.e., higher home posting increases the probability of toxicity at target), while negative values show "depolarizing" associations.

There are no universal associations based on whether the communities are on the same or opposing sides of the political spectrum; associations are unique to each pair. Furthermore, relationships are not necessarily reciprocal. For example, SOC users become less toxic on the DEM community as they post more at home, while the opposite holds for DEM users on SOC. In an oppositional pair example, increased echo chamber prevalence in TR makes toxicity more likely on SOC, while the opposite is true the other way around.

The largest polarizing relationship (EU-UK to SOC) amounts to a user posting close to 100% at home being about 2.5 times as likely to be toxic at the target compared to someone having posted nothing at home. Toxicity likelihood is drastically less likely (˜0.015 times) in the largest depolarizing effect (TR to LEFT). Again, this may be attributable to TR users being heavily moderated on LEFT. Overall, the directions and effect sizes vary heavily for unique community pairs. However, toxicity *associations* with increased engagement in the home community remain separate from the *actual* toxicity displayed by home users in another community. This is a distinction we clarify in the next section.

5 Typology of Community Relationships

Thus far, our analyses have focused on how political Reddit communities are related to one another in terms of their toxicity and echo chamber prevalence. Now, to better understand these cross-community dynamics, we synthesize our findings into a coherent typology based on three dimensions:

1. *Cross-community toxicity* (Figure 6). We define the relationship as:
   - inciting, if cross-toxicity ≥ 0.056 (highest quartile),
   - composed, if ≤ 0.031 (lowest quartile), and
   - basic, if $0.031 < toxicity < 0.056$.
2. *Increased engagement in the home community* (Figure 7). We define the relationship, with significance interpreted at $\alpha = 0.05$, as:
   - polarizing, if the pair model is significant and positive,
   - depolarizing if the model is significant and negative,
   - non-effectual if the model is non-significant (or cf, convergence failed).

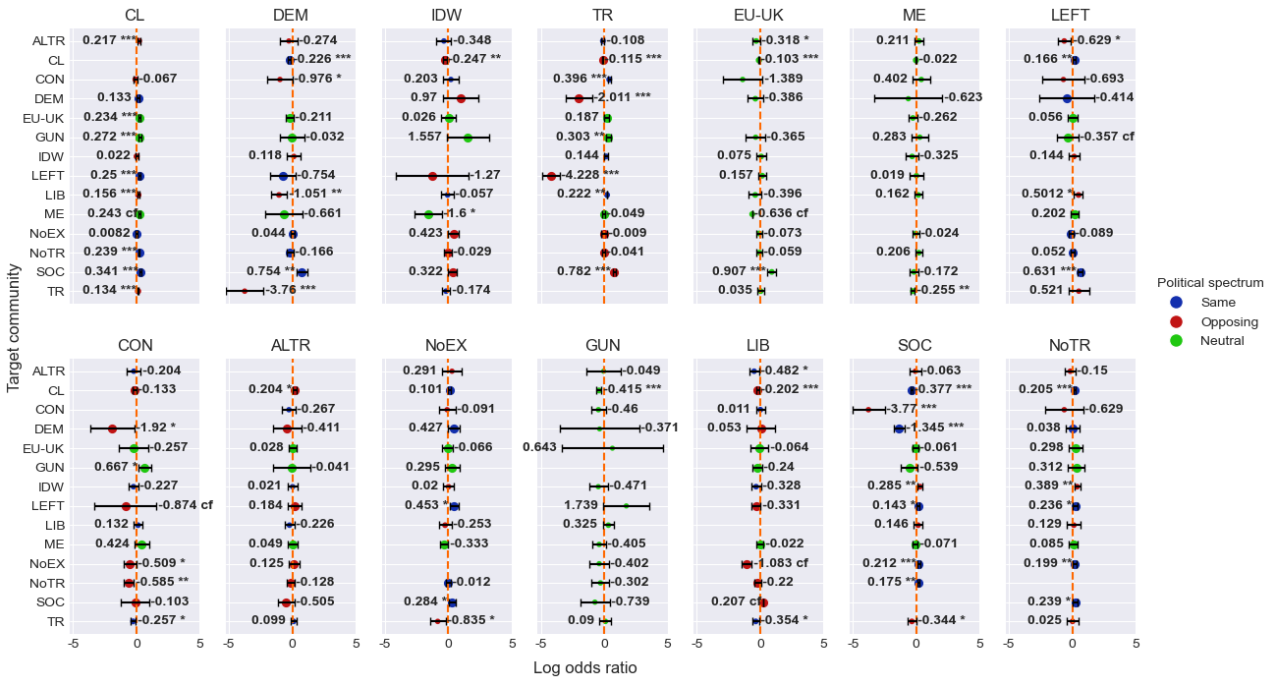

Figure 7: Forest-like plot showing pairwise regression estimates of echo chamber prevalence on toxicity probability at target community. $*p < 0.05$, $**p < 0.05/13$, $***p < 0.05/182$. cf = convergence failure. Error bars represent 95% confidence intervals. Community relation is same if both communities are right-leaning (IDW, TR, CON, ALTR, LIB) or both are left-leaning (CL, DEM, LEFT, NoEX, SOC, NoTR), opposing if one is left- and the other right-leaning, and neutral otherwise. Dot size is based on the number of observations relative to the largest number of observations. Notice the log-odds scale.

3. *Agreement in political leaning.* This is done based on a qualitative assessment of the communities' constituent subreddits, as discussed in Section 3. We define this as:
   - same, if both communities are right- or left-leaning,
   - opposing if one community is right- and the other left-leaning, and
   - neutral otherwise.

5.1 Typology Frequencies

Figure 8 is a mosaic plot showing the frequency of community relationship types. The most common type was an indifferent one (basic, non-effectual, and neutral) at 20.88% of the pairwise comparisons. By proportion, the basic and non-effectual types were more common among neutral pairs than opposing or same-spectrum pairs. However, basic and non-effectual was still the most common type within the subsets of opposing (e.g., LEFT to TR; LIB to SOC; ALTR to LEFT) and same-side (e.g., NoTR and DEM both ways; CON and LIB both ways) pairs.

Interestingly, inciting and polarizing types were more common in same-spectrum community pairs. Out of the 11 such same-side relationships observed, 9 occurred on the left, with the main "perpetrator" communities being SOC (to NoTR and NoEX), NoEX (to LEFT and SOC), and NoTR (to CL, SOC, LEFT, and NoEX). DEM was the only community that was neither an originator nor a receiver of this type on the left. The TR community was the sole originator of this type on the right (to LIB and CON). However, we also observe inciting and polarizing relationships with opposing-side pairs on four occasions (CL and ALTR both ways; SOC to IDW; NoTR to IDW). Additionally, depolarizing and composed relationships were most common among opposing pairs (e.g., DEM and CON both ways; Libertarian to CL) rather than same-side pairs; the only depolarizing and composed same-side pair was DEM to CL. This suggests that echo chamber-driven animosity may have predominantly occurred in politically agreeable communities.

Users were more likely to demonstrate hostility toward political outgroups when they interacted with ideologically aligned others (or within ideologically aligned communities) rather than when directly interacting with counter-partisans. To a lesser extent, ideologically aligned individuals also directed hostility toward each other due to "in-fighting" (see Section 5.2). However, as mentioned in Section 4.2, we stress that many of these patterns may be what *remained* on the communities following moderation, which leaves the possibility that toxic comments were selectively moderated based on the commenters' leaning.

We also observe some rare "wild card" types. Simultaneously inciting and depolarizing relationships (which only occurred with the opposing pairs TR to LEFT and NoEX to TR) suggest that there may exist learned civility amidst otherwise inciting discourse or that more frequent origin-community users may be more likely to have their comments

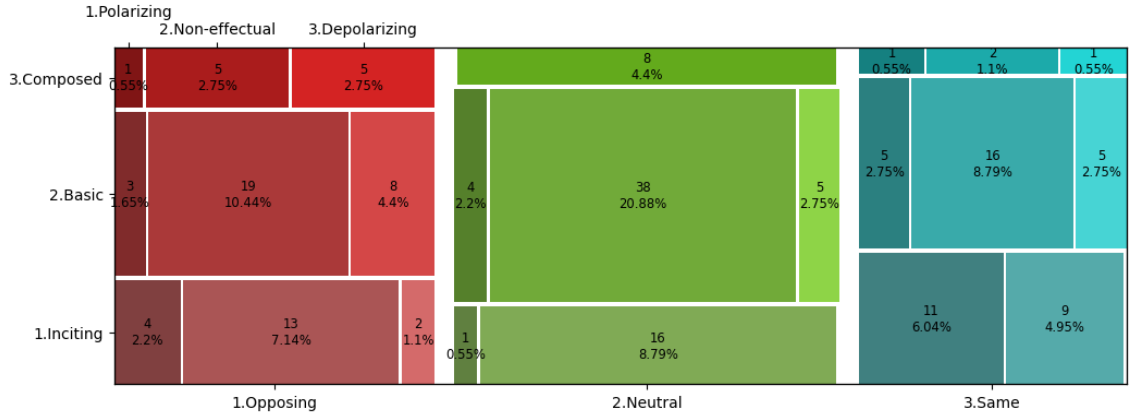

Figure 8: Mosaic plot showing number and proportions of typologized community relationships. Left axis: Type of speech based on cross-toxicity. Bottom axis: Political spectrum leaning. Top axis: Association with increased engagement at home.

removed on the target communities. Similarly, composed but polarizing relationships (DEM to SOC for same-side; LEFT to LIB for opposing) show that increased activity may lead to higher toxicity, even in otherwise civil discourse.

Overall, we find just one case of a depolarizing and composed relationship on the same side of the political spectrum, while we observe five such relationships for opposing-side pairs. At the same time, same-side pairs were more likely to have inciting and polarizing relationships, especially among left-wing communities (except for DEM, which was mostly involved in composed and depolarizing relationships). However, we also note that an inciting *and* polarizing relationship may not necessarily be more problematic than, say, *just* an inciting one. For example, ALTR, which was one of the most extreme communities in our dataset, was the originator of 7 inciting but non-effectual types; this means that ALTR users tended to be more toxic on many other communities, but they did not become *even more* toxic as they posted more at home. This could be due to, e.g., the ALTR community already being very high in toxicity, which would leave a smaller margin for an increase in toxicity levels.

5.2 Annotation Study

Next, we perform an annotation study to clarify whether the cross-toxicity among same-leaning communities points to political in-fighting between these communities or "ganging up" to collectively reprimand the political outgroup.

First, we create two data pools of toxic (as per Perspective API) left-to-left and right-to-right comments. Every comment in these pools is a top-level comment, i.e., a direct response to the submission. We do so as deeper-level comments lack the crucial context required to understand the comment's target.

We then randomly sample 400 comments, 200 from each pool. We extract the title, body (if any), hyperlinks (if any), and media such as images or videos (if any) included in the submission each comment is responding to for further context around the comment. Each comment is labeled by two annotators (two authors of this paper) as per two categories: toxicity (toxic or non-toxic) and target (outgroup-directed, ingroup-directed, unclear). For toxicity, we obtain good agreement (Krippendorff's α of 0.71 and 0.83 for left and right, respectively). For target, we obtain moderate α ratings of 0.59 and 0.52 for left and right, respectively. We resolve disagreements through a discussion of contested comments.

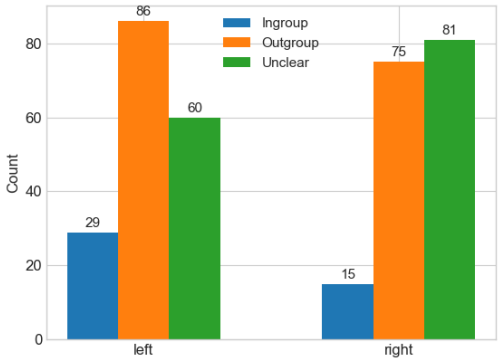

Figure 9: Targets of same-to-same leaning toxic comments as determined through annotations.

In Figure 9, we report the results of the annotation. Note that we omit misclassified non-toxic comments ($<$ 15% of the sample) as they do not fall in the study's aims. Ignoring comments with unclear targets, we find that the vast majority of toxic comments were indeed directed toward political outgroups, suggesting that the polarizing patterns we observe may be mostly due to same-leaning communities instigating rather than attacking each other. However, there was also a non-negligible proportion of comments (25.2% and 16.7% for left and right, respectively) demonstrating political in-fighting. This mostly reflects disagreements on endorsed politicians, issue positions, or clashes between ideologies (e.g., anarchism vs. state socialism).

Consistent with previous patterns, in-fighting was somewhat more frequent among the left. Although the annotation

study concerns a much smaller scale than our previous analyses, it confirms that polarization may occur mostly when same-leaning users interact with each other and speak negatively about political outgroups in those outgroups' absence.

6 Discussion & Conclusion

In this work, we present a large-scale, historical analysis of Reddit's political spaces between 2006 and 2019. We aim to determine whether the degree of engagement with echo chambers relates to behavior outside of them. We find that political communities on Reddit were more varied than the traditional left-right split during this period. Each community carried its norms both in the toxicity of conversations it hosted and in how its users behaved elsewhere.

Users predominantly engaged with their home communities but did show fairly diverse posting patterns. For **RQ1**, which concerns how echo chamber engagement relates to the probability of hostile intergroup interactions, we find that whether the degree of echo chamber (i.e., home) engagement related to toxicity in a target community depends on the unique relationship between the two communities. That is, increased or decreased polarization could occur both between and within political leanings, with these relationships not necessarily being reciprocal.

For **RQ2** on how the polarization and hostility relationships of communities vary based on political leaning, typologizing the communities revealed interesting patterns. Contrary to what one could expect, inciting and polarizing types were more common between communities lying on the same side of the political spectrum; however, this mostly reflected the reprimand of political outgroups rather than in-fighting. The presence of "wild card" combinations (e.g., polarizing and composed, inciting and depolarizing) suggests that political discourse is complex and influenced both by established cross-community and individual users' engagement patterns. Different communities had unique relationships, and echo chamber engagement did not act in unilateral directions. Nonetheless, content moderation possibly played a substantial (albeit unclear) role in the patterns observed.

6.1 Implications

Our work is a first attempt at bridging the echo chamber and hostile interaction perspectives of polarization, exploring how the two may be interdependent. Furthermore, it is among the first studies looking at the *degree* of echo chamber engagement at the user level rather than focusing on distinct chamber-like communities. The complex picture that arises from our study suggests that increased engagement with specific communities can broadly be associated with both polarization and depolarization of users. We also found more cross-polarization among left-wing communities, and this was mostly outgroup- (i.e., right-wing-) directed.

At the same time, we also observed more in-fighting among left-wing communities which could partly explain why left-wing radicalization is less common than right-wing radicalization (Hosseinmardi et al. 2020; Ribeiro et al. 2020) as left-wing users encounter more attitudinal disruption, and this may keep more extreme opinions in check. Arguably, our findings are important for several reasons.

High-level mapping. First, they situate user activity in the wider context of Reddit's political sphere. We utilize a dataset to capture a relatively complete set of *political* subreddits, which allows us to get a broader picture than would be possible by studying select subreddits over more restricted time periods. Thus, we can capture a high-level overview of the complex interplay between engagement patterns and toxic behavior, as well as highlight cases where diversifying user engagement could reduce hostile interactions (i.e., in polarizing relationships). A potentially fruitful direction to mitigate polarization could be to focus on organically occurring communities that ostensibly increase the diversity of engagement (e.g., *r/changemyview* or *r/AskTrumpSupporters*) and design systems that encourage and support more communities to form.

Indifferent communities. Second, we show that several communities within Reddit's political space were fairly neutral and indifferent toward each other in all respects; indeed, this was the most common type of relationship. Given that such communities hosted political discussions which were not particularly charged, they may be studied for their potential as "online buffer zones" where users' stances on individual issues, rather than their political leanings, are most salient. Overall, this pattern may suggest that political polarization is a more contextual rather than ubiquitous problem. However, we stress that, especially regarding cross-toxicity, our typology was *relative* to other relationships (i.e., only 25% of relationships could be treated as inciting due to quartile-based classification). At the same time, other work has found that political discussions, in general, tend to be more toxic than other kinds of conversations (Rajadesingan, Budak, and Resnick 2021); therefore, even our lower-toxicity relationships could be more toxic than relationships in domains other than politics. Further work is needed to verify how contextual polarization truly is.

Polarization in aligned communities. Third, while past work (Cinelli et al. 2021; Marchal 2022; Garimella et al. 2018) has focused on echo chambers and hostile interactions between counter-partisans as explanations for polarization, here we find inciting and polarizing patterns predominantly between politically aligned communities (especially among the left), but composed and depolarizing patterns predominantly between politically opposed ones. One potential explanation is that hostile interactions between opposing partisans may be more context-specific than previously thought, as they did not dominate when examining the wider community context. Another explanation is that toxic behavior may occur in real-time; however, this is retroactively moderated selectively only when this behavior comes from users whose views disagree with the wider community. Indeed, the likelihood of comment removals increases drastically when a community is negatively targeted by another on Reddit (Kumar et al. 2018). Regardless, polarization was observed largely in the form of agreeable communities inciting and reinforcing each other when speaking negatively about political outgroups.

These results could also carry important implications for content moderation. Particularly due to ideological biases or

subreddit-specific rules, users aligning with a community's political stance may be allowed to continue displaying toxic behavior as long as they do not cross partisan lines. In turn, this can result in evocative polarization even in the absence of ideological opponents. This is an important consideration that warrants future work, as it raises potential questions around the differences between stated and realized moderation goals (e.g., whether it is anti-hostility or anti-dissent).

6.2 Limitations and Future Work

Engagement vs. exposure. Although we start from the idea of exposure to diverse information, what we actually measure is commenting activity (i.e., engagement). However, many users may "lurk" in oppositional political spaces and view but never engage with posts. Therefore, we hope that future work will study the polarization phenomenon in the context of true information exposure, using metrics like clicks and reading time of different pieces of content (see, for example, Garimella et al. (2021)).

Hostility as toxicity. Our measure of toxicity may arguably only represent a small part of possible expressions of hostility; others include, for example, anger (Kumar et al. 2018) or inter-community attacks (Kumar et al. 2018; Datta and Adar 2019). Furthermore, this hostility may not necessarily be aimed toward the target community, but rather, a third outgroup community altogether. Future research could examine several such measures of hostility alongside each other (e.g., anger, toxicity, etc.) when studying these relationships to observe which expressions are the most dominant.

Non-causal inference. Our model uses time data to observe an effect (toxicity) following a previous event (home-community posting prevalence); however, this was not a truly causal effect as extraneous factors could be affecting both toxicity and posting prevalence. Future research could employ methods more suited to causal inference, such as controlled experiments or regression discontinuity analysis.

Content moderation. We aimed to study cross-toxicity between communities and whether this toxicity was more pronounced for users who demonstrated more one-sided engagement with their preferred communities. However, we did not clarify whether these patterns were due to moderation measures or naturally occurring. Polarized communities may have been more toxic due to more lax moderation, which could bias our results. Future research could distinguish between these two scenarios, as this is important for understanding how interactions of any type (i.e., intergroup or intragroup ones) arise online for other users to witness.

Selection of subreddits. We intentionally chose a wide range of subreddits to cluster based on the amount of political content they host to match the large scope of our research questions. However, in doing so, we also lost some qualitative information regarding these spaces. For example, An et al. (2019) only studied 4 subreddits, but these were carefully selected based on the specific political candidates they supported, whether contrarian discourse was allowed on the subreddit, and other unique characteristics.

Some of the specific inter-community relationships we observed might be due to the unique characteristics of these communities; for example, some may have predominantly hosted subreddits that advocated for specific political candidates, and others may have been pro- or anti-establishment, etc. Furthermore, the time span of our observation period was very large (13 years). Throughout this period, some users could have changed their political affiliations or issue positions. The long time span also opens the possibility that various events could have taken place that affected activity on Reddit and any communities which were active at the time but were not considered here.

While the scale of our analysis was a methodological choice to generalize our findings beyond specific cases, future research could adopt a more qualitative method of period and community selection to determine when and for which communities the different types of relationships hold. This is particularly important considering that the treatment of "echo chambers" in this study was relatively broad, and discourse within these bundled subreddits was likely more diverse and stemmed from differing levels of ideological heterogeneity than what would normally be expected in traditional echo chambers. We hope that future work can probe this within more ideologically homogeneous spaces.

Acknowledgments

We would like to thank Ashwin Rajadesingan for providing us with the initial list of 31K political subreddits. We also thank the anonymous reviewers for their valuable feedback.

This work was partially funded by the UK EPSRC grant EP/S022503/1, which supports the UCL Centre for Doctoral Training in Cybersecurity, the UK's National Research Centre on Privacy, Harm Reduction, and Adversarial Influence Online (REPHRAIN, UKRI grant: EP/V011189/1), and the US NSF under grants IIS-2046590, CNS-2114411, CNS-1942610, and CNS-2114407. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the funders.

References

An, J.; Kwak, H.; Posegga, O.; and Jungherr, A. 2019. Political Discussions in Homogeneous and Cross-Cutting Communication Spaces. In *ICWSM*.

Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. B. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *PNAS*, 115(37).

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239).

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10).

Bates, D.; Mächler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. In *ICWSM*.

Bliuc, A.-M.; Smith, L. G. E.; and Moynihan, T. 2020. "You wouldn't celebrate September 11": Testing online polarisation between opposing ideological camps on YouTube. *Group Processes & Intergroup Relations*, 23(6).

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10).

Brugnoli, E.; Cinelli, M.; Quattrociocchi, W.; and Scala, A. 2019. Recursive patterns in online echo chambers. *Scientific Reports*, 9(1).

Chaney, A. J. B.; Stewart, B. M.; and Engelhardt, B. E. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *RecSys*.

Cinelli, M.; Morales, G. D. F.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2021. The echo chamber effect on social media. *PNAS*, 118(9).

Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E*, 70(6).

Datta, S.; and Adar, E. 2019. Extracting Inter-Community Conflicts in Reddit. In *ICWSM*.

Datta, S.; Phelan, C.; and Adar, E. 2017. Identifying Misaligned Inter-Group Links and Communities. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW).

De Francisci Morales, G.; Monti, C.; and Starnini, M. 2021. No echo in the chambers of political interactions on Reddit. *Scientific Reports*, 11(1).

Del Vicario, M.; Gaito, S.; Quattrociocchi, W.; Zignani, M.; and Zollo, F. 2017. Public discourse and news consumption on online social media: A quantitative, cross-platform analysis of the Italian Referendum. arXiv:1702.06016. Accessed: 2020-10-21.

Dubois, E.; and Blank, G. 2018. The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5).

Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. In *WWW*.

Garimella, K.; Smith, T.; Weiss, R.; and West, R. 2021. Political Polarization in Online News Consumption. In *ICWSM*.

Gay, D. M. 1990. Usage Summary for Selected Optimization Routines. Technical report, AT&T Bell Laboratories.

Guess, A.; Nyhan, B.; Lyons, B.; and Reifler, J. 2018. Why selective exposure to like-minded political news is less prevalent than you think. Technical report, Knight Foundation.

Holtz, D.; Carterette, B.; Chandar, P.; Nazari, Z.; Cramer, H.; and Aral, S. 2020. The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify. In *ACM EC*.

Hosseini, H.; Kannan, S.; Zhang, B.; and Poovendran, R. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. arXiv:1702.08138. Accessed: 2022-05-13.

Hosseinmardi, H.; Ghasemian, A.; Clauset, A.; Rothschild, D. M.; Mobius, M.; and Watts, D. J. 2020. Evaluating the scale, growth, and origins of right-wing echo chambers on YouTube. arXiv:2011.12843. Accessed: 2020-11-27.

Jain, E.; Brown, S.; Chen, J.; Neaton, E.; Baidas, M.; Dong, Z.; Gu, H.; and Artan, N. S. 2018. Adversarial Text Generation for Google's Perspective API. In *CSCI*.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community Interaction and Conflict on the Web. In *WWW*.

Lees, J.; and Cikara, M. 2020. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*, 4(3).

Marchal, N. 2022. "Be Nice or Leave Me Alone": An Intergroup Perspective on Affective Polarization in Online Political Discussions. *Communication Research*, 49(3).

Matakos, A.; Aslay, C.; Galbrun, E.; and Gionis, A. 2022. Maximizing the Diversity of Exposure in a Social Network. *IEEE TKDE*, 34(9).

Mills, R. A. 2018. Pop-up political advocacy communities on reddit.com: SandersForPresident and The Donald. *AI & Society*, 33(1).

Moore-Berg, S. L.; Ankori-Karlinsky, L.-O.; Hameiri, B.; and Bruneau, E. 2020. Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *PNAS*, 117(26).

Nelder, J. A.; and Mead, R. 1965. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4).

Newman, M. E. J. 2006. Modularity and community structure in networks. *PNAS*, 103(23).

Pariser, E. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.

Rajadesingan, A.; Budak, C.; and Resnick, P. 2021. Political Discussion is Abundant in Non-political Subreddits (and Less Toxic). In *ICWSM*.

Rajadesingan, A.; Resnick, P.; and Budak, C. 2020. Quick, Community-Specific Learning: How Distinctive Toxicity Norms Are Maintained in Political Subreddits. In *ICWSM*.

Ribeiro, M. H.; Ottoni, R.; West, R.; Almeida, V. A. F.; and Meira, W. 2020. Auditing radicalization pathways on YouTube. In *ACM FAccT*.

Rogers, E. M.; and Bhowmik, D. K. 1970. Homophily-Heterophily: Relational concepts for communication research. *Public Opinion Quarterly*, 34(4).

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *ACL*.

Suhay, E.; Bello-Pardo, E.; and Maurer, B. 2018. The Polarizing Effects of Online Partisan Criticism: Evidence from Two Experiments. *The International Journal of Press/Politics*, 23(1).

Sunstein, C. R. 2001. *Republic.com*. Princeton University Press.

Terren, L.; and Borge-Bravo, R. 2021. Echo Chambers on Social Media: A Systematic Review of the Literature. *Review of Communication Research*, 9.

Vaccari, C.; Valeriani, A.; Barberá, P.; Jost, J. T.; Nagler, J.; and Tucker, J. A. 2016. Of Echo Chambers and Contrarian Clubs: Exposure to Political Disagreement Among German and Italian Users of Twitter. *Social Media + Society*, 2(3).

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing*, 17(4).

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

Zannettou, S.; Elsherief, M.; Belding, E.; Nilizadeh, S.; and Stringhini, G. 2020. Measuring and Characterizing Hate Speech on News Websites. In *ACM WebSci*.

Zhang, J. S.; Tan, C.; and Lv, Q. 2019. Intergroup Contact in the Wild: Characterizing Language Differences between Intergroup and Single-group Members in NBA-related Discussion Forums. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW).

Zhu, C.; Byrd, R. H.; Lu, P.; and Nocedal, J. 1997. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software*, 23(4).

Zollo, F.; Bessi, A.; Del Vicario, M.; Scala, A.; Caldarelli, G.; Shekhtman, L.; Havlin, S.; and Quattrociocchi, W. 2017. Debunking in a world of tribes. *PLOS ONE*, 12(7).