

Integrated or Segregated? User Behavior Change After Cross-Party Interactions on Reddit

Yan Xia,¹ Corrado Monti,² Barbara Keller,¹ Mikko Kivela¹

¹Department of Computer Science, Aalto University, Espoo, Finland

²CENTAI, Turin, Italy

yan.xia@aalto.fi, corrado.monti@centai.eu, barbara.keller@aalto.fi, mikko.kivela@aalto.fi

Abstract

It has been a widely shared concern that social media reinforces echo chambers of like-minded users and exacerbate political polarization. While fostering interactions across party lines is recognized as an important strategy to break echo chambers, there is a lack of empirical evidence on whether users will actually become more integrated or instead more segregated following such interactions on real social media platforms. We fill this gap by inspecting how users change their community engagement after receiving a cross-party reply in the U.S. politics discussion on Reddit. More specifically, we investigate if they increase their activity in communities of the opposing party, or in communities of their own party. We find that receiving a cross-party reply to a comment in a non-partisan discussion space is not significantly associated with increased out-party subreddit activity, unless the comment itself is already a reply to another comment. Meanwhile, receiving a cross-party reply is significantly associated with increased in-party subreddit activity, but the effect is comparable to that of receiving a same-party reply. Our results reveal a highly conditional depolarization effect following cross-party interactions in spurring out-party community engagement, which is likely part of a more general dynamic of feedback-boosted activity.

Introduction

Political polarization has been growing in the U.S. over the past few decades (Pew Research Center 2017; Iyengar et al. 2019; Finkel et al. 2020), both ideologically – characterized by increased opinion divide on ideological issues (McCarty, Poole, and Rosenthal 2016) – and affectively – characterized by increased animosity toward out-party members (Iyengar, Sood, and Lelkes 2012). The situation is alarming as political polarization can contribute to a harsh political climate (Jackson 2018), dehumanization (Harel, Jameson, and Maoz 2020), and even outbursts of violence (Michael 2017).

The rise of social media is seen as one of the factors that contribute to political polarization. In particular, researchers argue that on social media, algorithms create *filter bubbles* that isolate users from content they disagree with (Pariser 2011), while users themselves form *echo chambers* where they only receive information from and interact with like-minded others (Sunstein 2018).

To counter the formation and reinforcement of echo chambers on social media, researchers have proposed algorithms for fostering interactions between users of opposite political leanings (Garimella et al. 2017b; Musco, Musco, and Tsourakakis 2018; Garimella et al. 2017a; Tommasel, Rodriguez, and Godoy 2021). Real-world applications of similar strategies were reportedly taken into consideration by Twitter in 2018.¹

However, it is not clear whether these strategies will actually break echo chambers and bring polarized users together. While the intergroup contact theory (Allport, Clark, and Pettigrew 1954; Pettigrew 1998) implies that cross-party interactions may decrease polarization, and the results of many social experiments (Levendusky and Stecula 2021; Rossiter and Carlson 2024; Combs et al. 2023; de Jong 2024) align with this theory, other studies (Bail et al. 2018; Wojcieszak 2011) hint at the possibility that they may spark a *backfire effect* – that is, they may instead increase polarization and contribute to a further isolation of opposing sides. In addition to mixed evidence from experimental studies, there is a lack of empirical research on how real social media users behave after interacting with members of the opposing party.

What will fill this gap is an unobtrusive analysis of user behavior on a real social media site where cross-party interactions naturally happen. Reddit², one of the most popular news aggregation and discussion websites in the U.S., serves as an ideal platform of study, as researchers have found that the level of echo chamber effect is lower on Reddit than on Twitter and Facebook (Cinelli et al. 2021), and heterophilic interactions between members of opposing parties are even more common than homophilic interactions in Reddit discussions of U.S. politics (De Francisci Morales, Monti, and Starnini 2021; Monti et al. 2023).

We therefore set out to study the user behavior change following cross-party interactions in Reddit discussions of U.S. politics. Given the echo chamber narrative, our research revolves around the following question: after a user receives a reply from someone of the opposite political leaning, are they more likely to break out of their echo chamber and increase their activity in communities of the oppos-

¹<https://www.washingtonpost.com/technology/2018/08/15/jack-dorsey-says-hes-rethinking-core-how-twitter-works/>

²<https://www.reddit.com/>

ing party (i.e., *out-party activity*)? Or alternatively, are they more likely to increase their activity in communities of their own party (i.e., *in-party activity*)?

The temporal nature of Reddit data allows us to study the after-effect of cross-party interactions using a *before-observation-after* time frame. For an *observation period* of two months, we first collect all comments posted in *r/news*, the most popular non-partisan subreddit (i.e., subcommunity) that discusses U.S. news. Within the observation period, we then identify for each user the time they received the first reply in *r/news*, or, if they did not receive any, the time they posted the first comment in *r/news*. We consider the 30-day period before this time as the *before period* and the 30-day period after as the *after period*. After collecting the comment posting history of all involved users in the before period and the after period, we infer the party affiliation of each user based on their before period activity in a set of partisan subreddits we identified.

Using a regression model, we inspect the association between receiving a cross-party reply in *r/news* and participating more in out-/in-party subreddits in the after period than in the before period. We analyze the robustness of our results by running the model separately on data from each of five different years and observing how consistent the results are.

Our analysis shows that after a user receives a cross-party reply to a comment they posted in *r/news*, they are not necessarily more likely to increase their activity in out-party subreddits, unless the original comment is already a reply to another comment. Meanwhile, receiving a cross-party reply in *r/news* is significantly associated with subsequently increased in-party activity, but receiving a same-party reply is associated with a comparable effect.

Complementing previous experimental studies that measured the attitudinal or opinion change induced by cross-cutting conversation (Levendusky and Stecula 2021; Rossiter and Carlson 2024; Combs et al. 2023; de Jong 2024), we take a behavioral analysis perspective and reveal that in the U.S. politics discussion on Reddit, the depolarization effect of a single cross-party interaction in spurring the receiver's out-party community engagement is limited, unless the receiver already demonstrates a level of willingness to engage with other opinions; and in the latter case, the increase in out-party community engagement can be driven by a more general motivating effect of receiving peer feedback (Joyce and Kraut 2006; Burke, Marlow, and Lento 2009; Eckles, Kizilcec, and Bakshy 2016). However, we also see no strong evidence of a backfire effect, indicated by either a consistently more negative effect of a cross-party interaction in predicting increased out-party activity, or a consistently more positive effect of a cross-party interaction in predicting increased in-party activity, compared with that of a same-party interaction. Our study sheds new light on the dynamics of political polarization on social media platforms by showing the nuanced after-effects of cross-party communication in real-world settings.

Related Work

Political Polarization on Social Media. While there is mixed evidence on whether social media exacerbates polarization in offline settings (Allcott et al. 2020; Asimovic et al. 2021), online discussions on social media are shown to be extremely polarized. Empirical studies of Twitter data show that Twitter retweet networks are highly segregated on various topics across multiple countries (Conover et al. 2011; Morales et al. 2015; Garimella et al. 2016; Darwish 2019; Cossard et al. 2020; Salloum, Chen, and Kivelä 2022), indicating that retweeting – usually a sign of endorsement (Metaxas et al. 2015) – is largely skewed toward the same party. Similarly, researchers have found strong evidence of partisan sharing of information on Facebook (An, Quercia, and Crowcroft 2014; Schmidt et al. 2018).

The scarcity of cross-party endorsement hints at the existence of echo chambers on social media, where users are exposed to similar views and isolated from opposing views. However, some suggest that the echo chamber narrative might be exaggerated (Barberá 2020; Guess et al. 2018). For example, Barberá et al. (2015) showed that cross-party interactions take place more frequently than believed on Twitter – although these interactions can have a more negative tone (Mekacher, Falkenberg, and Baronchelli 2023). Barnidge (2017) found that more disagreement is perceived on social media than in other communication settings. Fletcher and Nielsen (2018) revealed that social media users consumed more diverse news sources than non-users. Flaxman, Goel, and Rao (2016) found the use of social networks and search engines to be associated with increased individual exposure to incongruent viewpoints, although it is also associated with increased ideological segregation between users: this challenges the argument that algorithms are what limit users' exposure to opposing views and thus exacerbate polarization. Bakshy, Messing, and Adamic (2015) further showed that individual selection plays a more important role than algorithms in limiting user exposure to cross-cutting content on Facebook, while Robertson et al. (2023) found that partisan news consumption on Google Search is driven primarily by user choices instead of algorithmic curation.

Polarization dynamics also differ significantly across platforms. For example, Cinelli et al. (2021) performed a comparative analysis of four well known social media platforms (Twitter, Facebook, Reddit, Gab) and found that Reddit embodied relatively smaller echo chamber effects than Facebook and Twitter in user interaction and information diffusion. Studies on Reddit more specifically showed the prevalence of cross-party interactions in Reddit political discussions (De Francisci Morales, Monti, and Starnini 2021), and found more evidence of demographic instead of ideological segregation among user interactions on Reddit (Monti et al. 2023).

Effect of Cross-Party Interactions. However, what is the effect of these cross-party interactions? Theory suggests that they may mitigate political polarization as intergroup contact reduces prejudice toward the outgroup (Allport, Clark, and Pettigrew 1954; Pettigrew 1998), but social experiments have shown mixed evidence regarding the effect of cross-

party exposure or conversations. Bail et al. (2018) found that exposure to cross-party news on social media can increase ideological polarization, while others found scarce evidence of such a backfire effect (also referred to as backlash or boomerang effect) (Guess and Coppock 2020; Wojcieszak et al. 2023; Bakshy, Messing, and Adamic 2015). A number of experimental studies on cross-party conversations suggest that they reduce affective polarization (Levendusky and Stecula 2021; Rossiter and Carlson 2024) but not ideological polarization (Levy 2021; de Jong 2024), while Combs et al. (2023) showed that anonymous cross-party conversations can decrease both affective and ideological polarization. Some researchers found that the depolarization effect takes place only when the conversation revolves around a casual topic (Santoro and Broockman 2022), or when the interlocutors are not strong partisans (Thomsen and Thomsen 2022). Some further showed that negative cross-party contact can instead exacerbate out-party hostility (Wojcieszak and Warner 2020).

Despite the considerable number of attempts to experimentally study the effect of cross-party interactions in synthetic settings, there has been limited empirical evidence on whether cross-party interactions depolarize real discussions on social media. Two survey studies before the rise of social media indicated that political diversity in one's social network may decrease one's participation in politics (Mutz 2002; Huckfeldt, Mendez, and Osborn 2004), and one empirical analysis of Facebook data echoes this finding by showing that political disagreement in one's Facebook network is associated with decreased turnout (Bond and Messing 2015). Most relevant to our work, Beam, Hutchens, and Hmielowski (2020) showed that exposure to cross-cutting news is related to affective depolarization on Facebook, while Heatherly, Lu, and Lee (2017) found that engagement in cross-cutting discussions on social networking sites is associated with less affective polarization. However, both these studies used self-reported survey data and were not able to reveal user dynamics in unobserved settings.

Using real user behavior data, early studies have shown that social interaction and feedback in general boosts user activity in online communities (Joyce and Kraut 2006; Burke, Marlow, and Lento 2009), while this finding is confirmed by an experimental study (Eckles, Kizilcec, and Bakshy 2016). However, the type of the interaction (i.e., cross-/same-party) presumably makes a substantial difference in its effect, as cross-party interactions are shown to be more negative (Marchal 2022) and more toxic (Wu and Resnick 2021) than same-party ones on social media. In our work, we focus specifically on user dynamics following cross-party interactions.

Method

We explore the effect of cross-party interactions by analyzing real user behavior data on Reddit. Reddit is a social media platform where users can create posts, comments, and replies of comments in different subreddits (i.e., subcommunities), while all the posts and comments can be voted up or down by other users.

We focus our analysis on the discussion of U.S. politics. On Reddit, the U.S. politics discussion is scattered across multiple subreddits, some of them being *partisan subreddits* that are dominated by users of the same political leaning (e.g. *r/The_Donald*, *r/hillaryclinton*), while others are *non-partisan subreddits* that engage users of different political leanings. A user's activity³ in partisan subreddits is often regarded an indicator of their political leaning (De Francis Morales, Monti, and Starnini 2021).

Considering the echo chamber narrative, we are interested in how users change their activity in partisan subreddits after a cross-party interaction in a non-partisan subreddit. Specifically, we pose the following research questions:

RQ1 After receiving a cross-party reply in a non-partisan subreddit, do users increase out-party subreddit activity (i.e., participate more in partisan subreddits of the opposite political leaning)?

RQ2 After receiving a cross-party reply in a non-partisan subreddit, do users increase in-party subreddit activity (i.e., participate more in partisan subreddits of the same political leaning)?

While we are interested in the after-effects of a cross-party reply compared with the case of no reply, we also compare the after-effects of a cross-party reply with those of a same-party reply, in order to recognize the part that can be explained by the general motivating effect of social interaction (Joyce and Kraut 2006; Burke, Marlow, and Lento 2009; Eckles, Kizilcec, and Bakshy 2016). We detail our data and methods as follows.

Data

We use data from the Pushshift Reddit data set (Baumgartner et al. 2020). We select *r/news* to be the non-partisan subreddit for sampling the cross-party interactions from, as it likely contains a large number of such interactions: *r/news* is consistently ranked among the top 15 most popular subreddits in previous years, and considered the largest subreddit that discusses U.S. news. Moreover, echoing the idea that news discussion is one of the main loci where opinions are formed and challenged (Kim, Wyatt, and Katz 1999), previous research identified that *r/news* accommodates both Democrats and Republicans, as well as interactions between them (Monti et al. 2023).

Considering the potential variation in dynamics across years, we collect data from five different years between 2014 and 2018. In order to have a feasible amount of data to analyze, for each year, we first collect all comments posted in *r/news* during a two-month *observation period*. We select August to September as the observation period because it is relatively close to the Election Day in the U.S., and thus in election years we expect there to be a considerable amount of discussion related to U.S. politics during this period. In the meantime, the period is not so close to the Election Day

³Throughout this paper, we use “activity”, “participation”, or “engagement” to refer to the act of posting comments that receive a *non-negative* score (i.e., number of upvotes subtracted by number of downvotes by other users) in a certain subreddit.

that the election results will play a major role in the discussion dynamics. Statistics of the collected data are reported in Table 1.

We extract the set of users who were active in this period, removing deleted accounts and automated accounts as detected by a previous Reddit study (Rollo et al. 2022). For each user, we pinpoint the time of their first received reply in `r/news` during the observation period; for those who did not receive any reply, we take the time of their first posted comment in `r/news` during the observation period. We consider 30 days before that as the *before period* of this user, and 30 days after as the *after period*. Then, we collect all comments they posted in *any* subreddit in the before period and the after period. As we will describe below, we use this information for characterizing the political leaning of each user, and for evaluating whether they increased their activity in partisan subreddits after an interaction in the observation period.

Leaning Detection

To identify cross-party interactions and partisan subreddit activity, we need to evaluate the leaning of each user and subreddit.

For this purpose, we first identify a list of *political subreddits*. To do so, we manually inspect for each year the top 1000 subreddits with the most comments in our data: for each subreddit, we scan through its first 5 posts and label it as a political subreddit if there is at least one post related to U.S. politics. With this approach, we find 87 political subreddits in total.

Then, we determine the leaning of each subreddit based on its political leaning score as given by Waller and Anderson (2021). In their paper, they devised a method for inferring the position of each subreddit along different social dimensions, by comparing the user base of the subreddit with that of certain seed subreddits with known positions (e.g., `r/democrats` and `r/Conservative` along the political dimension). Note that using this approach, the leaning score of each subreddit essentially reflects the leaning of its user base, as opposed to that of its topic.

We use a binned version of the subreddit leaning scores as done in the original paper: in this way, each subreddit s is associated with a political leaning score $\ell_s \in \{-2, -1, 0, 1, 2\}$, where -2 indicates extremely left-wing, -1 indicates left-wing, 0 indicates non-partisan, 1 indicates right-wing, and 2 indicates extremely right-wing. The list of political subreddits and their leaning scores can be found in the Appendix.

For each user u , we infer their leaning ℓ_u based on their activity in political subreddits during the *before period*. Specifically, we define an activity vector p_u for each user, where for each political subreddit s , we have its corresponding entry $p_{u,s}$ as the ratio of comments that user u posted in subreddit s over the total number of comments user u posted in all political subreddits during the before period. Note that we only count comments that received non-negative scores as valid activity: in this way, we exclude from the leaning calculation those comments that are negatively received by other users in the subreddit, which could indicate negative

behavior such as attacking or trolling, as opposed to proper engagement with the subreddit.

Then, the political leaning score ℓ_u of a user u is calculated as the weighted average of the subreddits leaning, i.e. $\ell_u = \sum_s p_{u,s} \cdot \ell_s$. We assign a Democratic leaning to a user if their leaning score is below 0, and a Republican leaning if their leaning score is above 0.

A *cross-party interaction* is then defined as a comment-reply interaction between two users who have opposite leanings, while a *same-party interaction* is a comment-reply interaction between two users of the same leaning. Similarly, out-party (resp. in-party) subreddit activity is defined as a user participating in a subreddit of the opposite (resp. same) leaning.

We create a regression data set where each row corresponds to one user and their first comment-reply interaction (if any) in `r/news` during the observation period. To ensure that both cross-/same-party interactions and out-/in-party subreddit activity are well-defined, we select the rows where the user has a non-zero inferred leaning; and for rows with an interaction, we selected those where both the user (i.e., comment author) and the replier (i.e., reply author) have non-zero inferred leanings.

To exclude irrelevant discussions, we limit the scope of our analysis to *sociopolitical* interactions. Following the definition by Moy and Gastil (2006), we consider a comment sociopolitical if it discusses at least one of the following entities or issues: (i) political figures, parties, or institutions; (ii) broad cultural and social issues (e.g., civil rights, moral values); (iii) national issues (e.g., healthcare, welfare). To select only comment-reply interactions where the reply is sociopolitical, we use the classifier developed by Monti et al. (2022) that was trained on a large Reddit data set and obtained an F1-score of 82% on a manually labeled subset. Statistics of the regression data set after filtering are presented in Table 1.

Average User Activity

To gain an initial understanding of user activity patterns before and after an interaction, we plot the average user activity in out-party subreddits for every two-day interval, from 30 days before the interaction to 30 days after (Figure 1), respectively for users receiving a cross-party reply and for those receiving a same-party reply.

We see that on average, out-party activity seems to have already increased before the interaction, presumably because of an activity peak around the time of the interaction. To provide a baseline of peaked activity, we additionally plot the average out-party activity of users receiving no reply, from 30 days before their first posted comment in `r/news` during the observation period to 30 days after. We find that compared with the baseline, the increase in out-party activity after an interaction seems to sustain for much longer, suggesting the existence of post-interaction dynamics that are not explained by peaked activity.

Regression Analysis

While the activity plot already reveals interesting patterns, it only presents compressed information from averaged statis-

	2014	2015	2016	2017	2018
# comments	822,237	1,156,138	1,154,495	1,621,039	1,424,752
# interactions	456,204	749,276	760,070	1,123,115	897,844
# accounts	121,886	164,292	163,798	211,307	220,338
# users	121,571	163,874	163,256	210,522	219,509
# non-zero-leaning users (left / right)	15,705 / 11,406	24,931 / 16,005	34,610 / 23,462	51,691 / 26,832	58,830 / 22,584
# users receiving reply	8,068	13,230	22,512	31,651	29,923
# users receiving sociopolitical reply	3,768	5,998	10,283	15,135	14,024
# users receiving cross-party reply (left / right)	917 / 941	1,407 / 1,434	2,316 / 2,551	3,811 / 3,711	3,177 / 3,204
# users receiving same-party reply (left / right)	1,292 / 618	2,281 / 876	4,016 / 1,400	6,002 / 1,611	6,585 / 1,058
# users receiving no reply (left / right)	2,250 / 1,849	3,315 / 2,300	4,852 / 3,867	7,717 / 4,190	9,486 / 3,608

Table 1: Statistics of the data set. The upper half shows the total number of comments in the observation period, the number of comment-reply interactions, the total number of accounts involved, the number of users excluding deleted and automated accounts, and the number of users with a non-zero inferred leaning. The lower half shows the statistics of our regression data set after filtering, including the number of users receiving a valid reply (i.e., both the user and the replier have a non-zero inferred leaning), those receiving a sociopolitical reply, those receiving a sociopolitical cross-party reply, those receiving a sociopolitical same-party reply, and those who received no reply during the observation period. For four of the rows, we report separately the number of users who are inferred to be left-/right-leaning.

tics. To utilize individual-level data when controlling for potential confounding factors, we perform a regression analysis as described below.

In order to evaluate whether users increase their activity in partisan subreddits after receiving a reply in $r/news$, we consider two dependent variables in our regression model:

- *More Out-Party Activity*: 1 if the user had more out-party subreddit activity in the after period than in the before period, and 0 otherwise.
- *More In-Party Activity*: 1 if the user had more in-party subreddit activity in the after period than in the before period, and 0 otherwise.

The key of the regression analysis is to compare the activity of users who received a cross-/same-party reply and the activity of those who did not receive any reply (i.e., baseline case), while controlling for other user features. Naturally, we have *Cross-Party Interaction* and *Same-Party Interaction* as the main independent variables in our regression model. Further, we speculate that there may exist a substantial difference between the comment-reply interactions where the comment is a top-level reply to some post (abbreviated as “top-level”), and those where the comment is a reply to some other comment (abbreviated as “nested”). In the latter case, especially if the comment replies to an out-party member, it is more likely that the user (i.e., comment author) already possesses a motivation to interact with the out-party, and thus may also increase activity in out-party subreddits. Therefore, we include four separate interaction variables in our regression model:

- *Cross-Party Interaction (Top-Level)*: 1 if the user received a cross-party reply to a comment that is a top-level reply to some post, and 0 otherwise.
- *Cross-Party Interaction (Nested)*: 1 if the user received a cross-party reply to a comment that is a reply to some other comment, and 0 otherwise.
- *Same-Party Interaction (Top-Level)*: 1 if the user received a same-party reply to a comment that is a top-level reply

to some post, and 0 otherwise.

- *Same-Party Interaction (Nested)*: 1 if the user received a same-party reply to a comment that is a reply to some other comment, and 0 otherwise.

Meanwhile, we introduce the following variables to control for the characteristics of the user as well as other features of the reply:

- User leaning:
 - *Left-Wing User*: 1 if the user has a leaning score below 0, and 0 otherwise.
 - *Hyper-Partisan User*: 1 if the user has a leaning score of -2 or 2, and 0 otherwise.
- User activity:
 - *User Out-Party Activity*: Number of comments posted by the user in subreddits of the opposite leaning in the before period, in logarithmic form, standardized by subtracting the mean and dividing by 2 standard deviations (an approach suggested by Gelman (2008)).
 - *User In-Party Activity*: Number of comments posted by the user in subreddits of the same leaning in the before period, in logarithmic form, standardized by subtracting the mean and dividing by 2 standard deviations.
 - *User Non-Partisan Activity*: Number of comments posted by the user in non-partisan subreddits in the before period, in logarithmic form, standardized by subtracting the mean and dividing by 2 standard deviations.
- Reply text (only for users who received a reply):
 - *Reply Length*: Number of characters in the reply, in logarithmic form, standardized by subtracting the mean and dividing by 2 standard deviations.
 - *Reply with URL*: 1 if the reply contains URL links, and 0 otherwise.

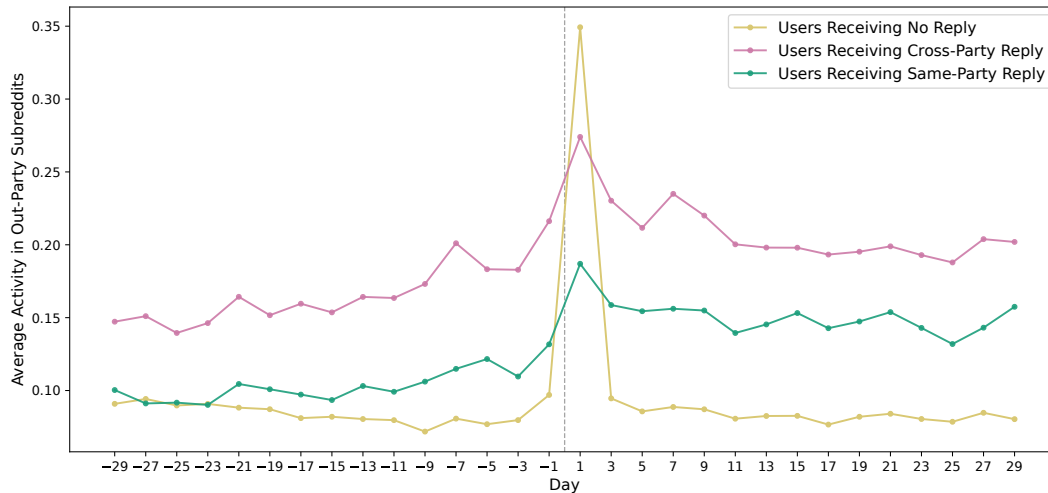


Figure 1: Average user activity in out-party subreddits for every two-day interval, from 30 days before to 30 days after receiving the first reply (resp. posting the first comment) in $r/news$ in the observation period, for users who received replies (resp. users who received no replies).

- *Negative Reply Text*: 1 if the VADER sentiment score (Hutto and Gilbert 2014) of the reply text is below 0, and 0 otherwise. The score falls in the range of $[-1, 1]$, where -1 indicates the most negative in sentiment, 1 indicates the most positive, and 0 indicates neutral.
- *Positive Reply Text*: 1 if the VADER sentiment score (Hutto and Gilbert 2014) of the reply text is above 0, and 0 otherwise.
- *Toxic Reply Text*: 1 if the toxicity score of the reply text returned by the Perspective API⁴ is above 0.7, and 0 otherwise. The score falls in the range of $[0, 1]$, where 0 indicates the least toxic, and 1 indicates the most toxic.
- Reply score (only for users who received a reply):
 - *Reply Score*: The user-voted score of the reply in logarithmic form, standardized by subtracting the mean and dividing by 2 standard deviations.
 - *Cross-Party Conflictual Scores*: 1 if the user and the replier have opposite leanings while the comment and the reply have scores of opposite signs, and 0 otherwise. Scores of opposite signs have been used in previous research as an indicator of a conflict between the user and the replier (Monti, De Francisci Morales, and Bonchi 2020).
 - *Same-Party Conflictual Scores*: 1 if the user and the replier have the same leaning while the comment and the reply have scores of opposite signs, and 0 otherwise.

Results

In this section, we first present the results of a model selection test, where we use the data aggregated from all studied

⁴<https://perspectiveapi.com>

years to select the most important independent variables for predicting the dependent variable. Then, we report the results of the regression analysis where we run a separate regression model on each year’s data, with the aim of revealing dynamics that are time-specific, and those that are consistent throughout time.

Model selection. To select the most important independent variables, we compare regression models with different sets of independent variables. Specifically, we gradually add groups of independent variables onto a baseline regression model and compare the Bayesian information criterion (BIC) (Schwarz 1978) values of the models before and after adding each group of variables. Since a lower BIC indicates a better model, we retain the group of independent variables if the BIC of the model decreases after adding the variables. Table 2 shows the models we tested and their difference from the baseline model in BIC, where the dependent variable is *More Out-Party Activity* for all models.

We start with baseline model *A* that includes all the user leaning and user activity variables. Upon adding the interaction variables, the BIC of the model decreases by 76 (model *B*), which indicates that the interaction variables substantially increase the model explanatory power even when considering a penalty for the increased number of variables. Therefore, we include the interaction variables in all following models.

By contrast, upon adding the reply text variables (model *C*), the model BIC increases, which means that these variables degrade the predictive power of the model, and thus we exclude them from our analysis. Similarly, none of the reply score variables (models *D* and *E*) is shown to be important to the model. Therefore, we choose model *B* (with only interaction, user leaning, and user activity variables) as the regression model for running our yearly analysis.

Note that we have also tried running the variable selection

	A	B	C	D	E
Δ BIC	0	-76	-30	-65	-55
Cross-Party Interaction (Top-Level)		0.080	0.057	0.080	0.078
Cross-Party Interaction (Nested)		0.177***	0.156***	0.177***	0.177***
Same-Party Interaction (Top-Level)		0.189***	0.164**	0.188***	0.157**
Same-Party Interaction (Nested)		0.262***	0.242***	0.262***	0.262***
Left-Wing User	-0.686***	-0.688***	-0.688***	-0.688***	-0.688***
Hyper-Partisan User	-0.318***	-0.241***	-0.241***	-0.241***	-0.241***
User Out-Party Activity	0.208***	0.213***	0.213***	0.213***	0.213***
User In-Party Activity	0.115***	0.108***	0.108***	0.108***	0.108***
User Non-Partisan Activity	0.259***	0.189***	0.189***	0.189***	0.189***
Reply Length			-0.021		
Reply with URL			-0.027		
Negative Reply Text			0.048		
Positive Reply Text			-0.014		
Toxic Reply Text			0.079		
Reply Score				0.001	
Cross-Party Conflictual Scores					0.006
Same-Party Conflictual Scores					0.106

Table 2: Results of the model selection test for choosing the most important independent variables, where we compare different regression models using aggregated data from all considered years. Each column (A, B, ...) corresponds to a regression model with a different set of independent variables, where the dependent variable is *More Out-Party Activity*. Δ BIC indicates the BIC score of the model compared with the baseline model *A*, where a lower BIC value indicates a better-performing model. Significance of each variable marked by ***: $p < 0.001$, **: $p < 0.01$, *: $p < 0.05$.

process on the other part of the data where the reply of the interaction is classified as non-sociopolitical. There, including the interaction variables increases the BIC of the model by 6, which indicates a decrease in its predictive power. This suggests that the correlations we find between interactions and subreddit activity in the sociopolitical counterpart of the data likely indicate real dynamics in political discussions, as opposed to artifacts of Reddit data that are present in both sociopolitical and non-sociopolitical interactions.

Aggregated analysis. Across all models with interaction variables, we see that receiving a nested cross-party interaction (i.e., receiving a reply to a comment that is already a reply to some other comment) seems significantly associated with more activity in out-party subreddits after the interaction ($p < 0.001$), while receiving a top-level cross-party interaction (i.e., receiving a reply to a comment that is a top-level reply to some post) demonstrates no significant association with more out-party activity afterward. This is in line with our speculation that users who receive a nested cross-party interaction might already possess a higher level of motivation to interact with the out-party, and thus are more likely to engage with out-party subreddits after the interaction. With this confounding effect excluded, our results indicate that a cross-party interaction alone is not significantly associated with an increase in subsequent out-party activity.

Interestingly, we observe that receiving a same-party interaction, top-level or nested, is significantly associated with

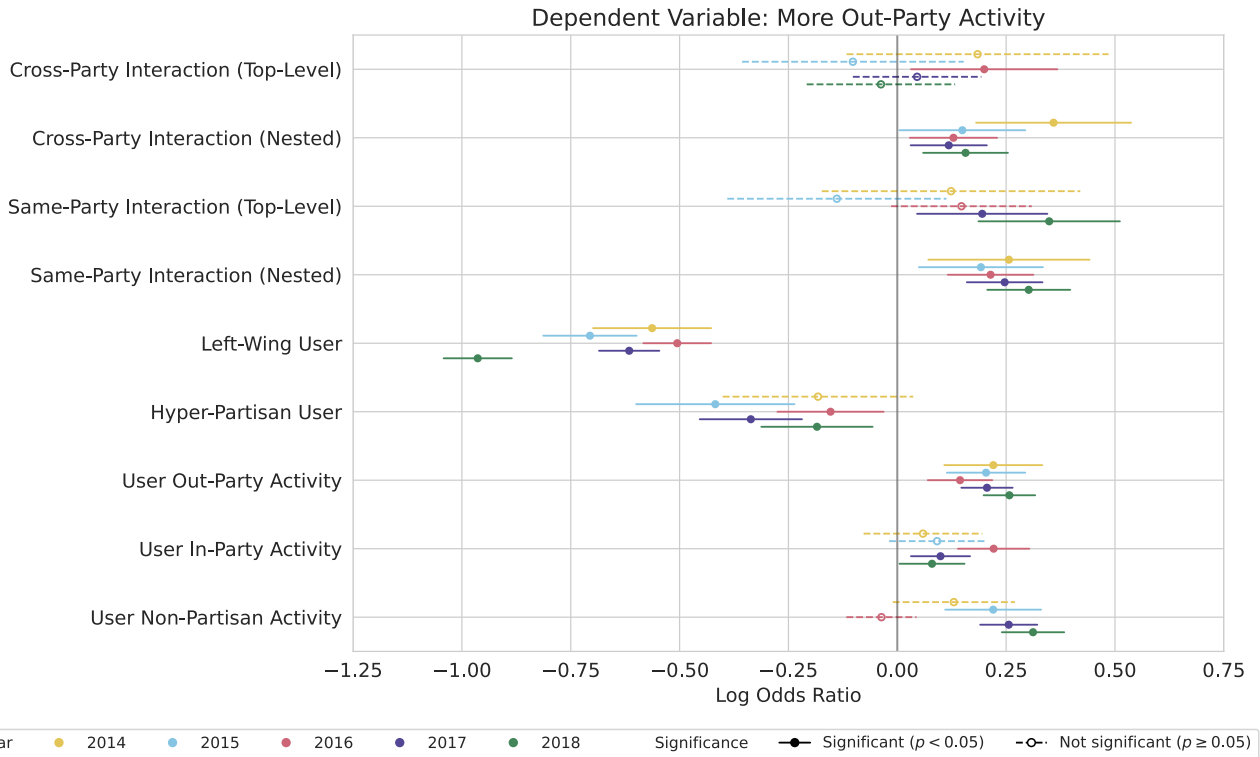
more activity in out-party subreddits after the interaction ($p < 0.001$ for both in the selected model *B*), while the coefficients are consistently larger than those in the corresponding cross-party case. We will further corroborate and analyze these results in the yearly analysis.

We also find that in all tested models, left-wing users are less likely to increase their activity in right-wing subreddits than vice versa ($p < 0.001$). This result aligns with previous findings on how right-wing users tend to be more involved in cross-cutting interactions than left-wing users (Wu and Resnick 2021; Eady et al. 2019; Grinberg et al. 2019). Similarly, we find that hyper-partisan users who dominantly participate in hyper-partisan subreddits are less likely to increase activity in out-party subreddits ($p < 0.001$), echoing previous studies that showed how hyper-partisan individuals are more difficult to depolarize (Wojcieszak 2011; Thomsen and Thomsen 2022; de Jong 2024).

Yearly analysis. We then run regression model *B* separately on data from each year between 2014 and 2018. The coefficients of each year’s regression model that predicts *More Out-Party Activity* are plotted in Figure 2A. We run the same regression model to predict the dependent variable *More In-Party Activity*, and plot the coefficients in Figure 2B.

We see that both nested interaction variables have significant positive correlations with both dependent variables in all years. This means that receiving a nested reply in *r/news*, whether cross-party or same-party, is consistently

A



B

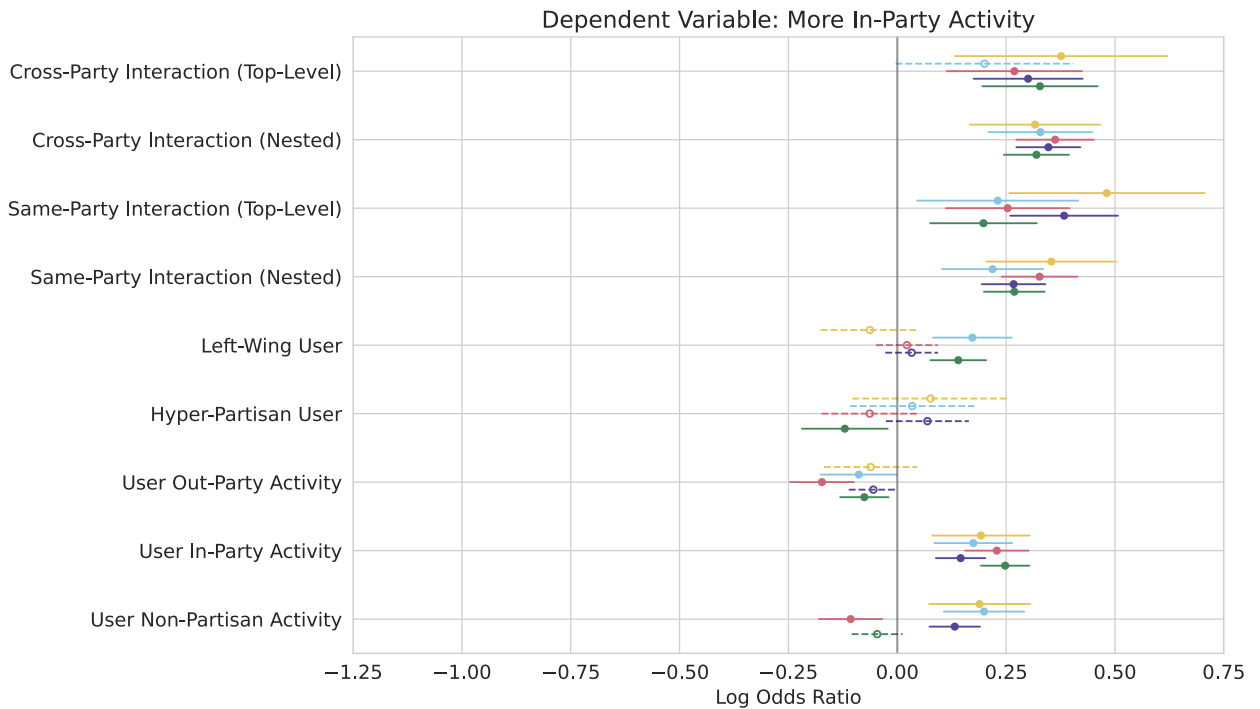


Figure 2: Coefficients (i.e., log odds ratio) of the regression model for predicting A) increased out-party subreddit activity, and B) increased in-party subreddit activity. Each row corresponds to an independent variable. Each point on the row corresponds to the estimated coefficient of the independent variable in one year's model, and the bar underneath indicates the 95% confidence interval of the estimate. A solid circle with a solid line indicates that the independent variable has a significant correlation with the dependent variable in the corresponding year, while a hollow circle with a dotted line indicates a non-significant one. A coefficient above 0 indicates that the independent variable has a positive correlation with the dependent variable, while a coefficient below 0 indicates a negative one.

associated with a user's increased activity in both out-party and in-party subreddits. This can be explained by how posting a comment as a reply to some other comment already reflects the user's elevated intention to actively engage with other opinions, which helps boost their subreddit activity when they receive a reply to the comment.

The coefficients of the top-level interaction variables demonstrate the correlation between receiving a reply and subreddit activity without the confounding effect in nested conversations. Figure 2A shows that receiving a reply to a top-level comment in *r/news*, whether cross-party or same-party, does not have a consistently significant association with increased out-party activity. However, the association is positive and significant for cross-party interaction in the election year 2016, and for same-party interaction in year 2017 and 2018, which might reflect some special discussion dynamics during the election period as well as after the new president took office. In general, the results do not indicate a strong depolarization effect following cross-party interactions, but we also see no strong evidence of a backfire effect, in which case a cross-party interaction is expected to be more negatively associated with increased out-party activity than a same-party interaction: Figure 3A shows that the association is not significantly more negative across all years (see Appendix for a more detailed discussion).

Meanwhile, in five (resp. four) out of five years, receiving a same-party (resp. cross-party) reply to a top-level comment in *r/news* is significantly associated with increased in-party subreddit activity (Figure 2B). This can be explained by the general motivating effect of social interaction in online communities (Joyce and Kraut 2006; Burke, Marlow, and Lento 2009; Eckles, Kizilcec, and Bakshy 2016). Here we do not observe a significantly larger effect of a cross-party interaction than that of a same-party interaction (Figure 3B), again finding no evidence of a backfire effect, in which case a cross-party interaction is expected to be more positively associated with increased in-party activity.

It is also interesting that the difference between the effect of a top-level interaction and that of a nested interaction is more pronounced in predicting increased out-party activity than in predicting increased in-party activity. We speculate that it is because out-party community engagement requires a much stronger motivation, as there exist substantial barriers toward intergroup contact (Ron et al. 2017). Therefore, the type of the interaction, which reflects the user's initial level of willingness to engage with other opinions, becomes a more important predictor of increased out-party activity.

Among other variables, we find confirmation that left-wing users are significantly less likely to increase activity in out-party subreddits (Figure 2A). Notably, the effect size of the *Left-Wing User* variable in predicting increased out-party activity is substantially larger than those of the interaction variables, indicating that user leaning probably plays a more critical role than social interactions in predicting out-party community engagement. Consistent with the results of our aggregated data analysis, hyper-partisan users are also significantly less likely to participate more in out-party subreddits (Figure 2A). However, user leaning seems to be a weak predictor of increased in-party activity.

In addition to our main regression model, we have also run an alternative model where the dependent variable is increased ratio of out-party subreddit activity. The results mostly echo our main findings (see Appendix for a more detailed discussion).

Discussion

Findings & Implications. We find that in the U.S. politics discussion on Reddit, receiving a cross-party reply to a posted comment in *r/news* is not necessarily associated with increased subsequent activity in out-party subreddits, unless the comment itself is already a reply to some other comment (RQ1). This indicates that the depolarization effect of a cross-party interaction in boosting engagement with out-party communities is highly conditioned on the receiver's intention to actively engage with the opinions of others. Even when there is a depolarization effect, it likely reflects a more general dynamic of feedback-boosted engagement that also explains a similar association between a same-party interaction and increased activity in out-party communities.

On the other hand, we do not see strong evidence that a cross-party interaction in a non-partisan community comes with a backfire effect of decreasing out-party activity or increasing in-party activity, beyond what can be explained by feedback-boosted engagement (RQ2). Compared with the predictive effect of a same-party interaction in our regression models, the effect of a cross-party interaction in predicting increased out-party activity is not consistently more negative, while the effect of a cross-party interaction in predicting increased in-party activity is not consistently more positive. This finding aligns with previous studies that found scarce backfire effects of exposure to counter-partisan information (Guess and Coppock 2020; Wojcieszak et al. 2023; Bakshy, Messing, and Adamic 2015).

Limitations. Our study used a limited data sample for the analysis. We ran the analysis on a single social media platform (i.e., Reddit), on a single community of interaction (i.e., *r/news*), within a data collection period of four months (i.e., July to October) across five different years (i.e., 2014-2018), and specifically in the U.S. politics context. While we look forward to future studies that replicate our findings under other settings, the results of our study are not guaranteed to generalize to other social media platforms, communities of interaction, time frames, or contexts. Also, we studied the effect following a single interaction, while multiple might be needed toward a more substantial change in user behavior.

For conducting a quantitative analysis, we had to operationalize theoretical constructs using observable user behavior logs on Reddit with some level of abstraction, and this process could have introduced inaccuracy into our conclusions. The user leaning inference in particular could have been disturbed by multiple sources of noise: we inferred a user's leaning based on their activity in partisan subreddits, where the partisan subreddits were identified based on a leaning score of each subreddit that only reflects the average leaning of its user base, and subreddit activity is measured

by the number of non-negative-score comments posted; we also worked with the assumption that users dominantly engage with in-party subreddits, which may not always hold. While we hope our current approach provides a good approximation for the corresponding theoretical constructs, our conclusions should be interpreted with this limitation in mind. More precise methods for detecting the political leaning of users, for example those that extract information also from the text they post, will help validate our results.

Finally, our study uses a regression analysis approach, which processes time series data in aggregated form. While such approach reveals user dynamics on a broader scale, future studies can deploy event history and time series analysis methods that make use of more fine-grained information. Moreover, regression analysis uncovers correlation instead of causality, therefore the effect we observed between cross-party interactions and subreddit activity could have been confounded by other factors. As much as we tried to control for confounding variables that are observable on Reddit (e.g., top-level or nested interaction, user leaning, user activity), we were not able to rule out the effect of unobservable confounders – for example, a user’s increased interest in the opposing party that stems from their real-life interactions may increase both their likelihood of receiving a cross-party reply in `r/news`, and their likelihood of increasing activity in out-party subreddits. Further evidence is needed to establish a causal link for the correlation we found. However, we believe that by presenting a correlation analysis, our work serves as an important first step toward casual studies that further answer the questions we posed.

Ethical Statement

We collected the publicly available Reddit data when the Reddit Data API granted free access as well as a worldwide, non-exclusive, non-transferable, non-sublicensable, and revocable license to all users (reddit 2018). Participants who generated the data agreed to the Reddit Privacy Policy (reddit 2014), and were aware of the public and freely accessible nature of the content they posted, as the subreddits are public, not password-protected, and participated by thousands of active subscribers. Reddit’s use of pseudonymous accounts reduces the likelihood of uncovering the true identity of users, and we further removed those pseudonyms in our preprocessing as a precaution. In our paper, we present only aggregated estimates without disclosing individual information.

To make our study reproducible, we host the preprocessed data and the code used for our analysis in a public Github repository⁵. While the original data may contain personally identifiable information or offensive content, we have removed all user identifiers in our preprocessed data, and converted the content of comments into linguistic feature scores.

There are potential broader impacts of our research. On the positive side, understanding the relationship between user interactions, user engagement, and political polarization in online discussions can inform the design of healthier social media platforms that alleviate polarization among

users. However, we acknowledge that our conclusions can be misinterpreted in an overly simplified and generalized way (e.g., “cross-party interactions on Reddit decrease political polarization”, “the content of a reply does not change its effect on the receiver”, etc.), potentially fueling inappropriate behavior in online discussions. Here we would like to reiterate that our conclusions are limited by the technical approach we have taken for our analysis, and should always be interpreted within the context of our study.

Acknowledgments

We thank the anonymous reviewers for their insightful comments on our manuscript. In addition, MK acknowledges funding from the Research Council of Finland (349366; 352561).

References

- Allcott, H.; Braghieri, L.; Eichmeyer, S.; and Gentzkow, M. 2020. The welfare effects of social media. *American economic review*, 110(3): 629–676.
- Allport, G. W.; Clark, K.; and Pettigrew, T. 1954. The nature of prejudice.
- An, J.; Quercia, D.; and Crowcroft, J. 2014. Partisan Sharing: Facebook Evidence and Societal Consequences. In *COSN’14: ACM Conference on Online Social Networks*, 13–24.
- Asimovic, N.; Nagler, J.; Bonneau, R.; and Tucker, J. A. 2021. Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences*, 118(25): e2022819118.
- Bail, C. A.; Argyle, L. P.; Brown, T. W.; Bumpus, J. P.; Chen, H.; Hunzaker, M. F.; Lee, J.; Mann, M.; Merhout, F.; and Volfovsky, A. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.
- Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239): 1130–1132.
- Barberá, P. 2020. Social media, echo chambers, and political polarization. *Social media and democracy: The state of the field, prospects for reform*, 34.
- 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): 1531–1542.
- Barnidge, M. 2017. Exposure to political disagreement in social media versus face-to-face and anonymous online settings. *Political communication*, 34(2): 302–321.
- Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435*.
- Beam, M. A.; Hutchens, M. J.; and Hmielowski, J. D. 2020. Facebook news and (de) polarization: Reinforcing spirals in the 2016 US election. In *Digital media, political polarization and challenges to democracy*, 26–44. Routledge.

⁵<https://github.com/yanxxia/cross-party-reddit>

- Bond, R.; and Messing, S. 2015. Quantifying social media's political space: Estimating ideology from publicly revealed preferences on Facebook. *American Political Science Review*, 109(1): 62–78.
- Burke, M.; Marlow, C.; and Lento, T. 2009. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 945–954.
- Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; and Starnini, M. 2021. The Echo Chamber Effect on Social Media. *PNAS: Proceedings of the National Academy of Sciences*, 118(9): e2023301118.
- Combs, A.; Tierney, G.; Guay, B.; Merhout, F.; Bail, C. A.; Hillygus, D. S.; and Volfovsky, A. 2023. Reducing political polarization in the United States with a mobile chat platform. *Nature human behaviour*, 1–8.
- Conover, M.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Menczer, F.; and Flammini, A. 2011. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, 89–96.
- Cossard, A.; De Francisci Morales, G.; Kalimeri, K.; Mejova, Y.; Paolotti, D.; and Starnini, M. 2020. Falling into the Echo Chamber: The Italian Vaccination Debate on Twitter. In *ICWSM '20: Fourteenth International AAAI Conference on Web and Social Media*, 130–140.
- Darwish, K. 2019. Quantifying polarization on twitter: the kavanaugh nomination. In *International conference on social informatics*, 188–201. Springer.
- De Francisci Morales, G.; Monti, C.; and Starnini, M. 2021. No Echo in the Chambers of Political Interactions on Reddit. *Scientific Reports*, 11: 2818.
- de Jong, J. F. 2024. Cross-partisan discussions reduced political polarization between UK voters, but less so when they disagreed. *Communications Psychology*, 2(1): 5.
- Eady, G.; Nagler, J.; Guess, A.; Zilinsky, J.; and Tucker, J. A. 2019. How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. *Sage Open*, 9(1): 2158244019832705.
- Eckles, D.; Kizilcec, R. F.; and Bakshy, E. 2016. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27): 7316–7322.
- Finkel, E. J.; Bail, C. A.; Cikara, M.; Ditto, P. H.; Iyengar, S.; Klar, S.; Mason, L.; McGrath, M. C.; Nyhan, B.; Rand, D. G.; et al. 2020. Political sectarianism in America. *Science*, 370(6516): 533–536.
- Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1): 298–320. Publisher: Oxford University Press tex.date-added: 2019-11-13 17:02:52 +0000 tex.date-modified: 2019-11-13 17:02:52 +0000.
- Fletcher, R.; and Nielsen, R. K. 2018. Are people incidentally exposed to news on social media? A comparative analysis. *New media & society*, 20(7): 2450–2468.
- Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2016. Quantifying Controversy in Social Media. In *WSDM '16: 9th ACM International Conference on Web Search and Data Mining*, 33–42.
- Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2017a. Factors in recommending contrarian content on social media. In *Proceedings of the 2017 ACM on Web Science Conference*, 263–266.
- Garimella, K.; De Francisci Morales, G.; Gionis, A.; and Mathioudakis, M. 2017b. Reducing controversy by connecting opposing views. In *Proceedings of the tenth ACM international conference on web search and data mining*, 81–90.
- Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine*, 27(15): 2865–2873.
- Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425): 374–378.
- Guess, A.; and Coppock, A. 2020. Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 50(4): 1497–1515.
- Guess, A.; Nyhan, B.; Lyons, B.; and Reifler, J. 2018. Avoiding the echo chamber about echo chambers. *Knight Foundation*.
- Harel, T. O.; Jameson, J. K.; and Maoz, I. 2020. The normalization of hatred: Identity, affective polarization, and dehumanization on Facebook in the context of intractable political conflict. *Social Media+ Society*, 6(2): 2056305120913983.
- Heatherly, K. A.; Lu, Y.; and Lee, J. K. 2017. Filtering out the other side? Cross-cutting and like-minded discussions on social networking sites. *New Media & Society*, 19(8): 1271–1289.
- Huckfeldt, R.; Mendez, J. M.; and Osborn, T. 2004. Disagreement, ambivalence, and engagement: The political consequences of heterogeneous networks. *Political Psychology*, 25(1): 65–95.
- Hutto, C. J.; and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Iyengar, S.; Lelkes, Y.; Levendusky, M.; Malhotra, N.; and Westwood, S. J. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22(1): 129–146.
- Iyengar, S.; Sood, G.; and Lelkes, Y. 2012. Affect, not ideology: a social identity perspective on polarization. *Public opinion quarterly*, 76(3): 405–431.
- Jackson, C. 2018. Fake News, Filter Bubbles, and Post-Truth Are Other People's Problems. *Global Market Research and Public Opinion Specialist (IPSOS)*, Washington DC.
- Joyce, E.; and Kraut, R. E. 2006. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3): 723–747.

- Kim, J.; Wyatt, R. O.; and Katz, E. 1999. News, talk, opinion, participation: The part played by conversation in deliberative democracy. *Political communication*, 16(4): 361–385.
- Levendusky, M. S.; and Stecula, D. A. 2021. *We need to talk: how cross-party dialogue reduces affective polarization*. Cambridge University Press.
- Levy, R. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3): 831–870.
- Marchal, N. 2022. “Be nice or leave me alone”: An intergroup perspective on affective polarization in online political discussions. *Communication Research*, 49(3): 376–398.
- McCarty, N.; Poole, K. T.; and Rosenthal, H. 2016. *Polarized America: The dance of ideology and unequal riches*. MIT Press.
- Mekacher, A.; Falkenberg, M.; and Baronchelli, A. 2023. The systemic impact of deplatforming on social media. *PNAS nexus*, 2(11): pgad346.
- Metaxas, P.; Mustafaraj, E.; Wong, K.; Zeng, L.; O’Keefe, M.; and Finn, S. 2015. What do retweets indicate? Results from user survey and meta-review of research. In *Proceedings of the international AAAI conference on web and social media*, volume 9, 658–661.
- Michael, G. 2017. The rise of the alt-right and the politics of polarization in America. *Skeptic (Altadena, CA)*, 22(2): 9–18.
- Monti, C.; Aiello, L. M.; De Francisci Morales, G.; and Bonchi, F. 2022. The language of opinion change on social media under the lens of communicative action. *Scientific Reports*, 12(1): 17920.
- Monti, C.; De Francisci Morales, G.; and Bonchi, F. 2020. Learning opinion dynamics from social traces. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 764–773.
- Monti, C.; D’Ignazi, J.; Starnini, M.; and De Francisci Morales, G. 2023. Evidence of Demographic rather than Ideological Segregation in News Discussion on Reddit. In *Proceedings of the ACM Web Conference 2023*, 2777–2786.
- Morales, A. J.; Borondo, J.; Losada, J. C.; and Benito, R. M. 2015. Measuring political polarization: Twitter shows the two sides of Venezuela. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(3): 033114.
- Moy, P.; and Gastil, J. 2006. Predicting deliberative conversation: The impact of discussion networks, media use, and political cognitions. *Political Communication*, 23(4): 443–460.
- Musco, C.; Musco, C.; and Tsourakakis, C. E. 2018. Minimizing polarization and disagreement in social networks. In *Proceedings of the 2018 world wide web conference*, 369–378.
- Mutz, D. C. 2002. The consequences of cross-cutting networks for political participation. *American Journal of Political Science*, 838–855.
- Pariser, E. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.
- Pettigrew, T. F. 1998. Intergroup contact theory. *Annual review of psychology*, 49(1): 65–85.
- Pew Research Center. 2017. The partisan divide on political values grows even wider. Technical report.
- reddit. 2014. reddit.com: privacy policy. <https://web.archive.org/web/20140101135359/http://www.reddit.com/help/privacypolicy>. Accessed on 2024-01-12.
- reddit. 2018. Reddit API Terms of Use. <https://web.archive.org/web/20180525145913/https://docs.google.com/a/reddit.com/forms/d/e/1FAIpQLSezNdDNK1-P8mspSbmtC2r86Ee9ZRbC66u929cG2GX0T9UMyw/viewform>. Accessed on 2024-01-12.
- Robertson, R. E.; Green, J.; Ruck, D. J.; Ognyanova, K.; Wilson, C.; and Lazer, D. 2023. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*, 618(7964): 342–348.
- Rollo, C.; De Francisci Morales, G.; Monti, C.; and Panisson, A. 2022. Communities, Gateways, and Bridges: Measuring Attention Flow in the Reddit Political Sphere. In *International Conference on Social Informatics*, 3–19. Springer.
- Ron, Y.; Solomon, J.; Halperin, E.; and Saguy, T. 2017. Willingness to engage in intergroup contact: A multilevel approach. *Peace and Conflict: Journal of Peace Psychology*, 23(3): 210.
- Rossiter, E. L.; and Carlson, T. N. 2024. Cross-partisan conversation reduced affective polarization for republicans and democrats even after the contentious 2020 election. *The Journal of Politics*, 86(4): 1608–1612.
- Salloum, A.; Chen, T. H. Y.; and Kivelä, M. 2022. Separating polarization from noise: comparison and normalization of structural polarization measures. *Proceedings of the ACM on human-computer interaction*, 6(CSCW): 1–33.
- Santoro, E.; and Broockman, D. E. 2022. The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science advances*, 8(25): eabn5515.
- Schmidt, A. L.; Zollo, F.; Scala, A.; Betsch, C.; and Quattrociochi, W. 2018. Polarization of the vaccination debate on Facebook. *Vaccine*, 36(25): 3606–3612.
- Schwarz, G. 1978. Estimating the dimension of a model. *The annals of statistics*, 461–464.
- Sunstein, C. 2018. *# Republic: Divided democracy in the age of social media*. Princeton university press.
- Thomsen, J. P. F.; and Thomsen, A. H. 2022. Intergroup contact reduces affective polarization but not among strong party identifiers. *Scandinavian Political Studies*.
- Tommasel, A.; Rodriguez, J. M.; and Godoy, D. 2021. I want to break free! Recommending friends from outside the echo chamber. In *Proceedings of the 15th ACM Conference on Recommender Systems*, 23–33.
- Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888): 264–268.

Wojcieszak, M. 2011. Deliberation and attitude polarization. *Journal of Communication*, 61(4): 596–617.

Wojcieszak, M.; de Leeuw, S.; Menchen-Trevino, E.; Lee, S.; Huang-Isherwood, K. M.; and Weeks, B. 2023. No polarization from partisan news: Over-time evidence from trace data. *The International Journal of Press/Politics*, 28(3): 601–626.

Wojcieszak, M.; and Warner, B. R. 2020. Can interparty contact reduce affective polarization? A systematic test of different forms of intergroup contact. *Political Communication*, 37(6): 789–811.

Wu, S.; and Resnick, P. 2021. Cross-Partisan Discussions on YouTube: Conservatives Talk to Liberals but Liberals Don't Talk to Conservatives. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 808–819.

Ethics Checklist

1. For most authors...
 - (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? **Yes, see the Ethical Statement section.**
 - (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, see the Limitations subsection in the Discussion section.**
 - (e) Did you describe the limitations of your work? **Yes, see the Limitations subsection in the Discussion section.**
 - (f) Did you discuss any potential negative societal impacts of your work? **Yes, see the Ethical Statement section.**
 - (g) Did you discuss any potential misuse of your work? **Yes, see the Ethical Statement section.**
 - (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, see the Ethical Statement section.**
 - (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes.**
2. Additionally, if your study involves hypotheses testing...
 - (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
 - (b) Have you provided justifications for all theoretical results? **NA**
 - (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**
 - (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**
 - (e) Did you address potential biases or limitations in your theoretical framework? **NA**
 - (f) Have you related your theoretical results to the existing literature in social science? **NA**
 - (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**
3. Additionally, if you are including theoretical proofs...
 - (a) Did you state the full set of assumptions of all theoretical results? **NA**
 - (b) Did you include complete proofs of all theoretical results? **NA**
4. Additionally, if you ran machine learning experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **NA**
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **NA**
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **NA**
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **NA**
 - (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **NA**
 - (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? **NA**
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
 - (a) If your work uses existing assets, did you cite the creators? **Yes, see the Data subsection in the Method section.**
 - (b) Did you mention the license of the assets? **Yes, see the Ethical Statement section.**
 - (c) Did you include any new assets in the supplemental material or as a URL? **No, we only publish our code and preprocessed data in a github repository.**
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **Yes, see the Ethical Statement section.**
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes, see the Ethical Statement section.**
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **NA**
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **NA**
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
 - (a) Did you include the full text of instructions given to participants and screenshots? **NA**
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? **NA**
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **NA**
 - (d) Did you discuss how data is stored, shared, and de-identified? **NA**

Appendix A: Additional Regression Results

Difference-in-effects models. In addition to our main regression model that shows the absolute effects of cross-party interactions and same-party interactions, we fit another *difference-in-effects* model to inspect the differences in their effects. In this alternative model, we replace the *Same-Party Interaction* variables with *Interaction* variables, so that the coefficient of a *Cross-Party Interaction* variable indicates the difference between the effect of a cross-party interaction and that of an interaction (either cross-party or same-party).

The results are shown in Figure 3. Overall, we do not see a significant difference in effects that is consistent across all inspected years. While there does seem to be a growing trend of the difference over time where a cross-party interaction more and more negatively predicts increased out-party activity than an average interaction, it is unclear if we can interpret it as a growing backfire effect, as the growth in difference is mainly driven by an increasingly positive association between a same-party interaction and increased out-party activity, rather than an increasingly negative association between a cross-party interaction and increased out-party activity (Figure 2).

Ratio models. We fit another regression model to inspect the association between an interaction and increased ratio of out-party activity in the after period, as opposed to increased absolute activity that we inspect in the main model. More specifically, we have *More Out-Party Activity Ratio* as the dependent variable: let a_o denote out-party subreddit activity and a_i denote in-party subreddit activity, the dependent variable takes value 1 if the $a_o/(a_o + a_i)$ ratio is higher in the after period than in the before period, and 0 otherwise.

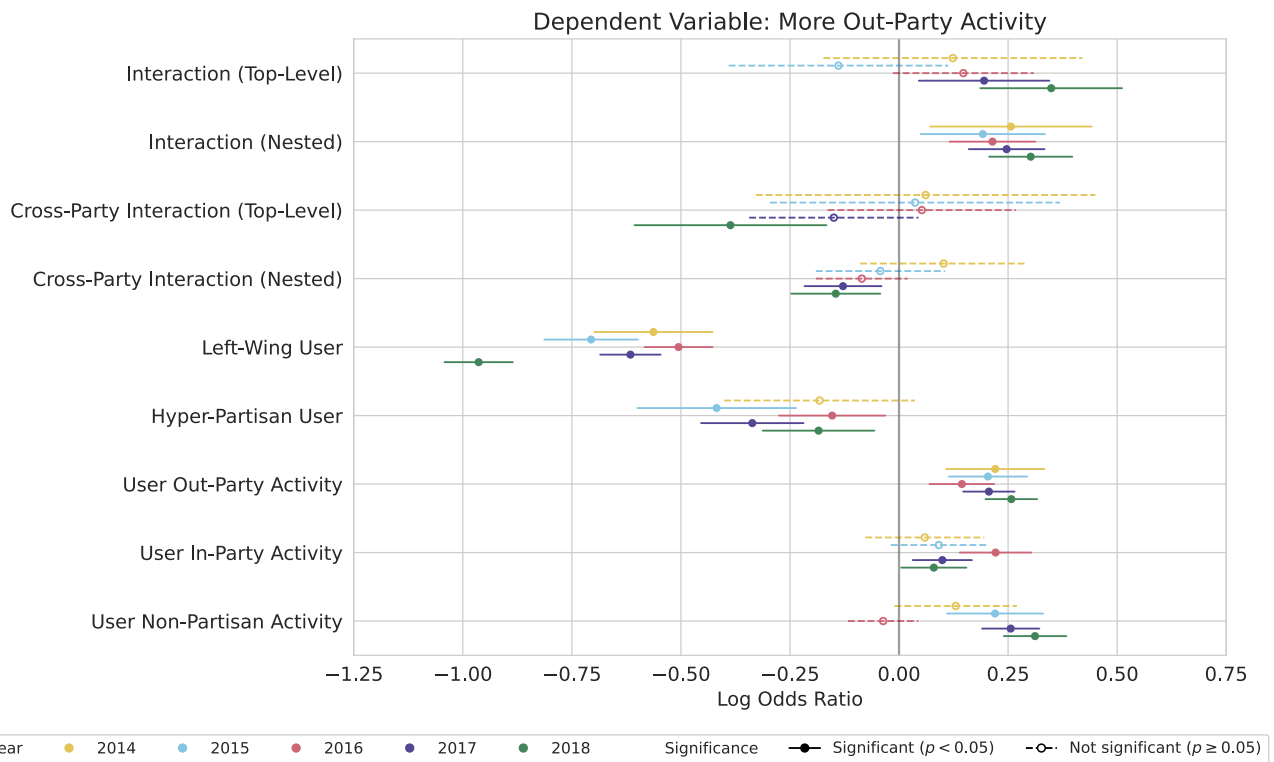
The results of respectively the absolute-effect variant (where we have independent variables of *Same-Party Interaction* and *Cross-Party Interaction*) and the difference-in-effects variant (where we have independent variables of *Interaction* and *Cross-Party Interaction*) of this ratio model are shown in Figure 4. We mostly see similar patterns as those in our main results: following a cross-party interaction, there is no clear evidence of either a depolarization effect that increases out-party activity ratio, or a backfire effect that decreases out-party activity ratio (Figure 4A). Again, a cross-party interaction seems to be more and more negatively predicting increased out-party activity ratio over time than an average interaction (Figure 4B), but the growing difference comes from an increasingly positive association between a same-party interaction and increased out-party activity ratio.

Note that for fitting this ratio model, we chose to exclude users who had no subreddit activity in the after period to avoid zero-division issues (while they all had nonzero subreddit activity in the before period because they had a valid leaning). After the filtering, the remaining number of users is 5,209 out of 7,867 in 2014, 7,996 out of 11,613 in 2015, 14,203 out of 19,002 in 2016, 20,677 out of 27,042 in 2017, and 20,758 out of 27,118 in 2018. Our findings are thus specific to this subset of users, and are not guaranteed to hold upon inclusion of zero-activity users via alternative approaches.

Appendix B: Political Subreddits

Table 3, 4, and 5 list respectively the left-leaning subreddits, non-partisan subreddits, and right-leaning subreddits we identified, along with their corresponding leaning scores.

A



B

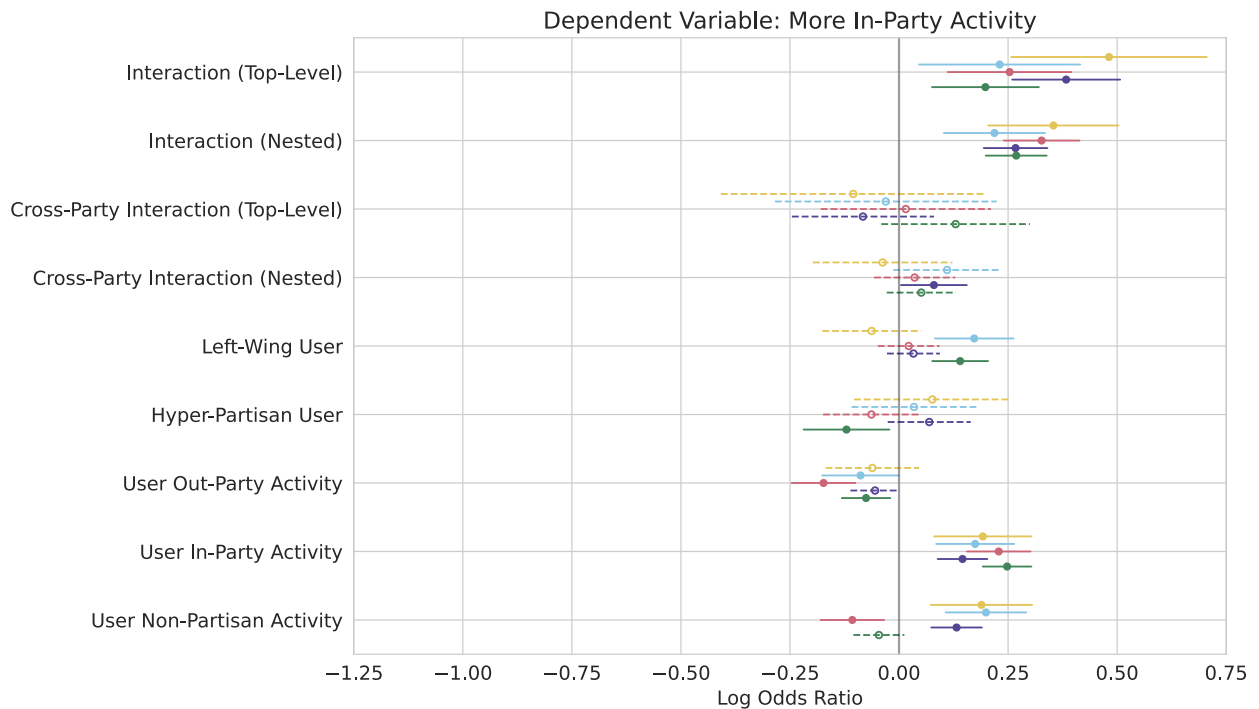
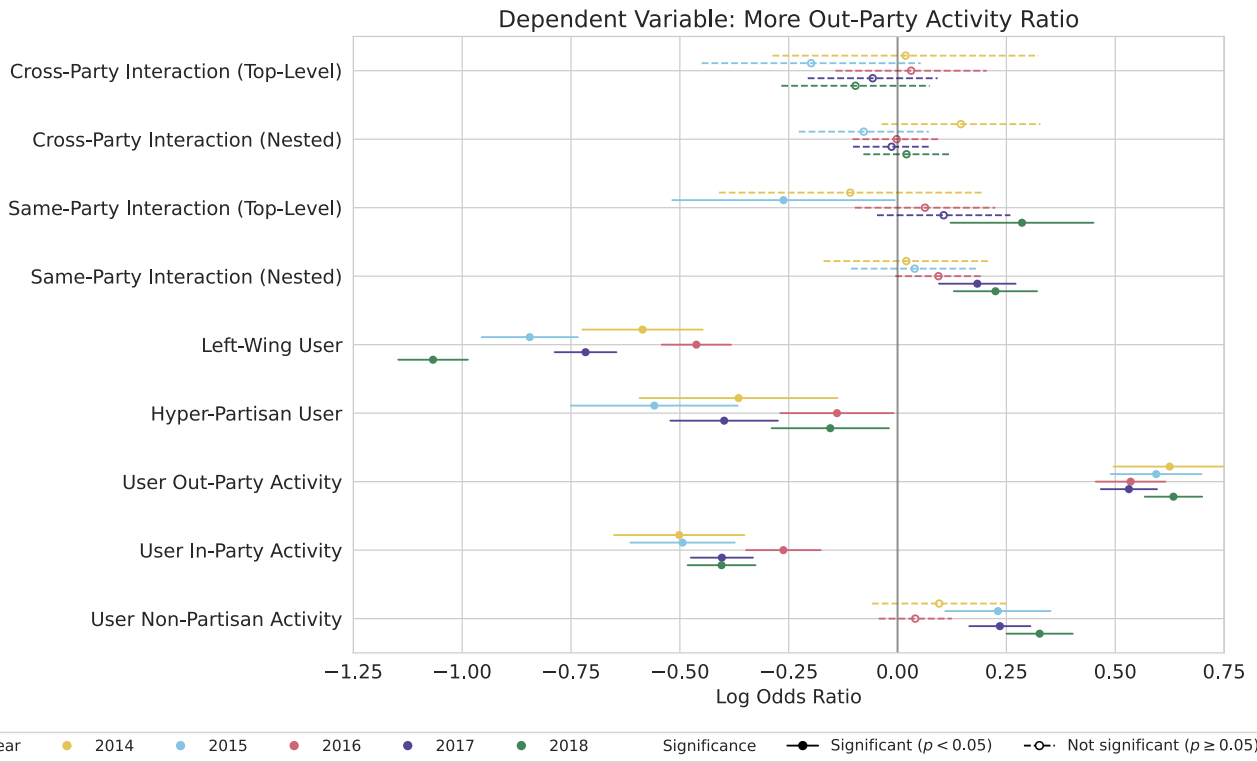


Figure 3: Coefficients (i.e., log odds ratio) of the difference-in-effects regression model for predicting A) increased out-party subreddit activity, and B) increased in-party subreddit activity, where the coefficient of a *Cross-Party Interaction* variable indicates the difference between the effect of a cross-party interaction and that of an interaction (either cross-party or same-party). Each row corresponds to an independent variable. Each point on the row corresponds to the estimated coefficient of the independent variable in one year's model, and the bar underneath indicates the 95% confidence interval of the estimate. A solid circle with a solid line indicates that the independent variable has a significant correlation with the dependent variable in the corresponding year, while a hollow circle with a dotted line indicates a non-significant one. A coefficient above 0 indicates that the independent variable has a positive correlation with the dependent variable, while a coefficient below 0 indicates a negative one.

A



B

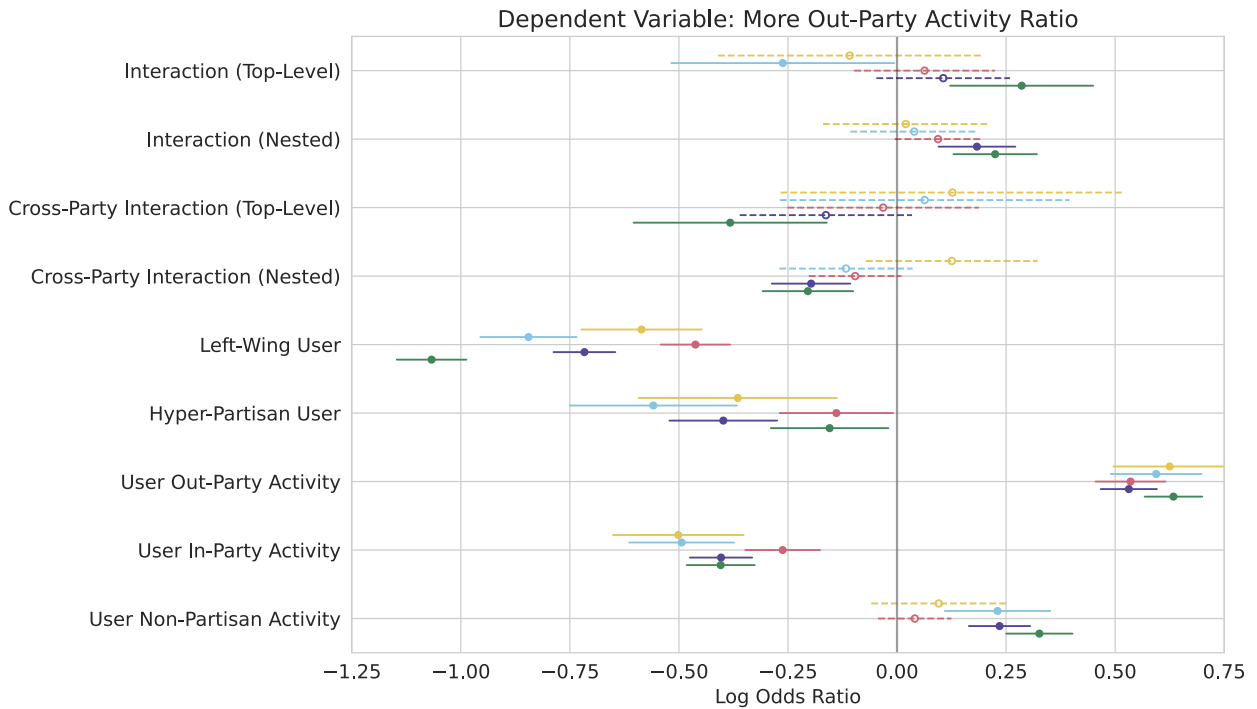


Figure 4: Coefficients (i.e., log odds ratio) of the A) absolute-effect regression model, and B) difference-in-effects regression model for predicting increased ratio of out-party subreddit activity. Each row corresponds to an independent variable. Each point on the row corresponds to the estimated coefficient of the independent variable in one year's model, and the bar underneath indicates the 95% confidence interval of the estimate. A solid circle with a solid line indicates that the independent variable has a significant correlation with the dependent variable in the corresponding year, while a hollow circle with a dotted line indicates a non-significant one. A coefficient above 0 indicates that the independent variable has a positive correlation with the dependent variable, while a coefficient below 0 indicates a negative one.

Subreddit	Leaning Score
r/democrats	-2
r/EnoughLibertarianSpam	-2
r/hillaryclinton	-2
r/progressive	-2
r/BlueMidterm2018	-2
r/Enough_Sanders_Spam	-2
r/GunsAreCool	-2
r/EnoughTrumpSpam	-2
r/Impeach_Trump	-2
r/Trumpgret	-2
r/LateStageCapitalism	-2
r/FULLCOMMUNISM	-2
r/Fuckthealtright	-2
r/ShitLiberalsSay	-2
r/COMPLETEANARCHY	-2
r/ChapoTrapHouse	-2
r/Political_Revolution	-2
r/Liberal	-2
r/asianamerican	-2
r/conspiratard	-2
r/liberalgunowners	-2
r/socialism	-2
r/The_Mueller	-2
r/Kossacks_for_Sanders	-2
r/SandersForPresident	-2
r/BasicIncome	-2
r/TrumpCriticizesTrump	-1
r/neoliberal	-1
r/askhillarysupporters	-1
r/BannedFromThe_Donald	-1
r/WayOfTheBern	-1
r/MarchAgainstTrump	-1
r/jillstein	-1
r/Anarchism	-1
r/The_Dotard	-1
r/atheism	-1
r/PoliticalHumor	-1
r/inthenews	-1
r/politics	-1

Table 3: List of left-leaning subreddits.

Subreddit	Leaning Score
r/Ask.Politics	0
r/AnythingGoesNews	0
r/PoliticalDiscussion	0
r/WhitePeopleTwitter	0
r/Bad.Cop.No.Donut	0
r/moderatepolitics	0
r/AskALiberal	0
r/DNCleaks	0
r/law	0
r/Economics	0
r/POLITIC	0
r/WikiLeaks	0
r/economy	0
r/news	0
r/PoliticalVideo	0
r/NeutralPolitics	0
r/JusticeServed	0
r/GaryJohnson	0
r/CapitalismVSocialism	0
r/ShitRConservativeSays	0

Table 4: List of non-partisan subreddits.

Subreddit	Leaning Score
r/Conservative	2
r/The_Donald	2
r/conservatives	2
r/uncensorednews	2
r/AskTrumpSupporters	2
r/altright	2
r/AskThe_Donald	2
r/ShitPoliticsSays	2
r/CCW	2
r/Catholicism	2
r/Republican	2
r/progun	2
r/Firearms	2
r/climateskeptics	2
r/guns	2
r/Anarcho.Capitalism	2
r/CringeAnarchy	2
r/Libertarian	2
r/TumblrInAction	2
r/HillaryForPrison	2
r/Christianity	2
r/MensRights	1
r/DebateAltRight	1
r/GoldandBlack	1
r/POTUSWatch	1
r/gunpolitics	1
r/ProtectAndServe	1
r/conspiracy	1

Table 5: List of right-leaning subreddits.