Social Network Analysis and Mining (2024) 14:22
https://doi.org/10.1007/s13278-023-01184-8

ORIGINAL ARTICLE

# Do users adopt extremist beliefs from exposure to hate subreddits?

**Matheus Schmitz[1] · Goran Muric[1] · Daniel Hickey[2] · Keith Burghardt[1]**

Received: 27 March 2023 / Revised: 19 August 2023 / Accepted: 11 December 2023


**Abstract**
Social media offers an avenue for like-minded individuals to interact in ways that were previously not possible. Yet, it can also be a breeding ground for hate and extremism to spread. Despite research into hate speech on social media, its influence on users adopting extremist beliefs is less understood. In this study, we use causal analysis to quantify extremist adoption resulting from users becoming active in hate online communities known as *subreddits*. Using an interrupted time series research design, we compare users who became involved in hate subreddits (treatment group) to those who did not (control group). This analysis is reproduced across ten different subreddits covering four different topics: Alt-Right, Racism, Sexism, and Fat-Shaming. From these analyses, we uncover a causal link between a user becoming active in a hate community and using more hate speech both within hate subreddits and across the wider platform. The results are consistent and replicate across communities. Our findings are tentative evidence that users adopt extremist ideas from exposure to hate subreddits.

**Keywords** Hate speech · Reddit · Social media · Interrupted time series · Regression discontinuity design

## 1 Introduction

Social media platforms must juggle between two important but conflicting goals: preserving freedom of speech and curbing hate speech and extremism. By swinging the pendulum too much toward curbing undesired ideas, platforms may become monotonous, ultimately losing their appeal to the users. This paper explores the converse case when a social media platform, Reddit (https://www.reddit.com), gave communities too much leeway to promulgate their ideologies, resulting in the formation of highly toxic hate groups. We seek to understand the impact that those problematic groups had on the dissemination of hate through the platform as a whole by exploring the following research question:

*How does becoming active in a hate subreddit change a user's behavior on the platform?*

The reason we explore this question is to better understand whether (and how) users adopt extremist ideas from exposure to hate subreddits, where we define "exposure" as posting messages on that subreddit. We answer this question by matching treatment users, i.e., those who become active in a hate subreddit, with control users, who are similar in behavior but never join said hate subreddit. We then model users' hate speech through interrupted time series design, to obtain a causal effect estimate of how much change in hate speech can be attributed as a result of becoming active in a hate subreddit. In total, we analyzed ten subreddits across 4 different categories: Alt-Right, Racist, Sexist, and Fat-Shaming. In contrast to a previous paper (Schmitz et al. 2022), we use a novel method to systematically label hate subreddits and use this larger pool of subreddits to find statistically significant effects of joining distinct hate subreddits and subreddit categories.

For each subreddit, we use sparse additive generative models of text (SAGE) (Eisenstein et al. 2011) to generate a lexicon of the community-specific hate slurs. The lexicons

✉ Keith Burghardt
keithab@isi.edu

Matheus Schmitz
mschmitz@isi.edu

Goran Muric
gmuric@isi.edu

Daniel Hickey
hickeyda@oregonstate.edu

[1] Information Sciences Institute, University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292, USA

[2] Department of Botany and Plant Pathology, Oregon State University, 2701 SW Campus Way, Corvallis, OR 97331, USA

Published online: 11 January 2024

are then used to measure the hate speech of each treatment user through time, both before and after they join the hate community, as well as measuring their matching control user over the same time period. Causal analysis through interrupted time series leads us to conclude that becoming active in a hate community leads to a measurable spillover in hate speech to other non-hate communities, meaning that a user's hate is not self-contained within the subreddit that they join. The findings are consistent across subreddits and categories, which demonstrates the qualitative and quantitative robustness of this effect. Our results provide tentative evidence that users adopt extremist ideas (Marwick et al. 2022), and point to the need for social media platforms to improve their moderation in order to reduce hate both within and outside hate communities.

## 2 Related work

### 2.1 Online communities

Motivations for joining online communities can be neutral or benign, such as boredom, seeking social support, and information exchange (Ridings and Gefen 2004; Laeeq 2017; Lewis et al. 2018; Yao et al. 2021; Stockdale Laura and Coyne Sarah 2020). Yet, motivations can also be malicious, such as attacking the community and its members (Jhaver et al. 2018), subsequently diminishing overall community activity (Kumar et al. 2018).

The continued engagement of users with the community hinges on several factors, such as whether a newcomer receives reciprocal interaction from existing community members (Arguello et al. 2006; Burke et al. 2009; Burke and Settles 2011; Kraut and Resnick 2012), with engagement being further boosted when replies are positive (Arguello et al. 2006) or personalized (Kraut and Resnick 2012). New editors making their first contributions to Wikipedia are less likely to continue attempting to contribute if their original edits are reverted (Halfaker et al. 2011; Zhang and Zhu 2006). For this reason, Wikipedia encourages existing users to be gentle with newcomers (Contributors Wikipedia 2022). The expression of negative emotion by other group members leads to community abandonment (Yao et al. 2021). A user's stay in a community is also related to the user's adoption of the group's linguistic norms (Danescu-Niculescu-Mizil et al. 2013). At times, certain newcomers are undesired (Kraut and Resnick 2012; Choi et al. 2008), such as those who do not conform to community ideology and culture (Boero and Pascoe 2012; Santos et al. May 2020).

The characteristics of each community on Reddit are sufficiently idiosyncratic that they can be accurately identified on both their style (86.5%) or topic (71.1%) (Tran and Ostendorf 2016). Rudeness and attacks between dissimilar people reduce online engagement across ideological lines (Marchal 2020), yet content about the out-group is shared at higher rates than content regarding the in-group (Rathje et al. 2021).

### 2.2 Hate speech in social media

It has been shown that news-feed algorithms favor users who share hate and extremist content over those whose posts are neutral or positive (Phadke and Mitra 2021). Further, Facebook has been shown to play a larger role in the active recruitment of new members, whereas Twitter is more prominent in terms of maximizing the broadcasting of the groups' ideologies (Phadke and Mitra 2020).

Physical characteristics, such as ethnicity, height, weight, and gender, alongside observable behavior, such as sensitivity or insecurity, are the primary attributes used by attackers in their hate speech (Mondal et al. 2017; Silva et al. 2016). When the hate speech is directed, as opposed to generalized, it tends to be angrier (ElSherief et al. 2018), and an increase in time spent on far-right websites such as Gab has been linked to an increase in hate speech (Gallacher and Bright 2021; Zannettou et al. 2018). When it comes to social media, it is estimated that as much as 25% of content contains some form of hate speech, with explicit hate speech being observed in 13.7% and implicit hate speech in 15.5% (Rieger et al. 2021). Hate users display changes in both activity levels and lexicon compared to regular users (Chatzakou et al. 2017).

Extremist subreddits have been shown to prop extremist content through user-level actions such as upvotes and downvotes that favor an echo chamber of surfacing content aligned with the groups' behavior while dismissing contrarian views (Gaudette et al. 2021; Massanari 2017; Gothard 2020). One important challenge in studying hate speech is the inherent difficulty of appeasing all critics with regard to the definition used, while also construing a definition that is specific enough to be implemented as software in a programmatic matter (Silva et al. 2016). Users who receive replies are less likely to become engaged in hate subreddits than users who do not, while the opposite effect is observed for non-hate subreddits, and this effect is attributable to the toxic, negative, and attacking nature of replies in hate subreddits (Hickey et al. 2023).

### 2.3 Moderation of hate speech

Both soft approaches such as quarantines, where users receive a warning and must agree to proceed before viewing the contents of toxic communities, as well as hard approaches such as banning problematic subreddits, have been shown to systemically reduce hate speech on Reddit (Chandrasekharan et al. 2017, 2022; Copland 2020). There

**Table 1** Subreddits studied and their category

| Subreddit | Category | Total users | Selected users | Bandwidth (days) |
|---|---|---|---|---|
| r/frenworld | Alt-right | 5916 | 2805 | 360 |
| r/honkler | Alt-right | 3321 | 740 | 360 |
| r/milliondollarextreme | Alt-right | 8509 | 5113 | 35 |
| r/CoonTown | Racist | 9681 | 5109 | 35 |
| r/GreatApes | Racist | 3141 | 2493 | 80 |
| r/WhiteRights | Racist | 5006 | 2849 | 360 |
| r/Braincels | Sexist | 4732 | 2260 | 360 |
| r/Incels | Sexist | 20,210 | 11,315 | 345 |
| r/MGTOW | Sexist | 6473 | 3966 | 30 |
| r/fatpeoplehate | Fat-shaming | 10,241 | 7349 | 360 |

is, however, evidence that such actions merely forward the hate users from Reddit to other less regulated platforms (Copland 2020). Put together, these indicate that there is likely a user overlap between different hate subreddits and their external counterparts, such that any moderation actions taken must account for user migration (Zannettou et al. 2017). For example, it has been shown that the banning of a subreddit leads its most active users to reduce their overall activity levels the most, but most users did not see a reduction in overall activity levels, rather they had only a reduction in their usage of subreddit specific language (Trujillo et al. 2021). The reduction in moderation on Twitter that followed Elon Musk's acquisition has led to a significant increase in hate speech on the platform (Hickey et al. 2023). Another challenge to moderation is that suppression of hate content in one platform can inadvertently boost the sharing of that content in another platform (Johnson Neil et al. 2019). Both as a result of moderation initiatives, as well as due to a myriad of other changing factors, it has been observed that certain communities can migrate, sometimes with the intent of avoiding moderation, and sometimes for other reasons (Davies et al. 2021). A factor that helps Reddit maintain users is its long tail of niche content, which can only be lively due to the size of Reddit's user base. This indicates that whether a hate or fringe community can be exterminated by a Reddit ban is related to the community's prominence and its ability to attract new members when residing in less popular platforms (Newell et al. 2016).

## 3 Methods

### 3.1 Identifying communities of interest

We base our selection of hate subreddits on prior work by Hickey et al. We previously identified 25 hate subreddits that are now banned (Hickey et al. 2023), from which we subsampled ten subreddits across the Racism, Sexism, Alt-Right, and Fat-Shaming categories, as shown in Table 1. The code used to parse and analyze data in our paper is available.[1]

### 3.2 Identifying subreddit-specific hate language

Our approach to defining hate words follows Supreme Court Justice Potter Stewart's principle for defining a threshold for obscenity: "I know it when I see it." Namely, words that are observed to have hate connotations in their contexts are considered hate words. This lexicon-based approach allows us to implement our analysis in a computationally-compatible manner.

We use sparse additive generative models of text (SAGE) (Eisenstein et al. 2011) to rank all unigrams from the studied subreddit with regard to how distinctive (i.e. community-specific) they are in relation to the typical word usage on Reddit. To define typical word usage, we gather a 10GB corpus containing posts (submissions and comments) randomly sampled across all of Reddit. For a given studied subreddit, we use all posts within it as the community corpus. We remove stopwords from both corpora. Many hate words are slang that are not well handled by stemming or lemmatizing, hence we do not apply any such techniques.

We further filter the ranked word list of each subreddit to the top 100 most distinctive words, and then rate them using a 3-rater system, where authors of this paper independently score the word as 0, 1, or 2, respectively, meaning *not a hate word*, *sometimes a hate word*, or *Always a hate word*. Words with a cumulative score at or above 4 (i.e., at least one rater rated the word as "always a hate word") are then defined as subreddit-specific hate words. Creating

[1] Code to analyze data and reproduce results: https://github.com/Matheus-Schmitz/Reddit_Hate/.

a customized lexicon for each hate subreddit is important because much of the hate expressed in subreddits are based on group-specific slang that is poorly captured by generalized lexicons (van der Does et al. 2022; Gerrard 2018). From these hate word lists, we can then quantify the hate speech for a given user on a given day as the ratio of hate words to total words. The complete hate lexicon for each subreddit is available in our GitHub page.[2]

## 3.3 Data gathering

In our study "treatment" is defined as becoming active in a hate subreddit. Users who become active are considered "treated users" while users who never become active are "control users". We explain how we gather each set of users below.

### 3.3.1 Treatment data

For each user who posted on each hate subreddit, we collect their entire posting history using the PushShift API (Baumgartner et al. 2020). To prevent our analysis from being influenced by automated accounts (bots), we manually inspect accounts whose name matches the keywords "bot", "auto", "transcriber", "gif", "link", "twitter", removing those that are self-identified as bots or are likely bots based on their behavior. Similarly, for each subreddit we manually inspect the top 20 accounts by activity levels, again removing those found to be automated. Overall, a small number of accounts ($< 1\%$) are removed.

To increase signal, we further filter treatment users to include only those who, at the time of becoming active in their respective hate subreddit, are not a part of any of the 25 other hate subreddits compiled by Hickey et al. (2023). We follow that procedure under the rationale that users who are already adopting extremist ideas as a result of interactions on other hate subreddits are unlikely to show a significant change in behavior when becoming active in 'just one more' hate subreddit. Table 1 displays both the original number of users collected, as well as the number of users being considered after filtering.

### 3.3.2 Control data

Our goal is to match a treated user to its most similar control user prior to the users becoming active in a hate subreddit. We thus opt for a targeted crawling approach, focusing on obtaining candidate control users that are most likely to be as similar as possible to the treated users. To that end, we obtain our control users from the top thirty subreddits with the highest percentage of user base overlap with each hate subreddit, always removing from that list users who are also active on the treatment subreddit itself, as those are considered treatment users.

### 3.3.3 Banned subreddits

We obtain a non-exhaustive list of banned subreddits to analyze hate speech by users within under-moderated communities outside the hate subreddit. Reddit does not divulge its bans, and hence all information on bans comes from user-generated content and self-reports. The gathering process was entirely manual and consisted of browsing Reddit itself for posts compiling subreddit bans. Given the nature of the data-gathering process, prominent bans and bans of large subreddits are more likely to be featured in our set. In total, we obtained 3515 subreddits reported to be banned.

### 3.3.4 Categorical aggregation

One negative side effect of filtering our treatment data to consider only users with no prior activity in other hate subreddits is that it reduces the overall sample size, which can make it hard to detect whether the measured effects are statistically significant or not. For this reason, we pool subreddit data in four categorical groups, as shown in Table 1. Since results are similar both across individual subreddits and grouped categories, we opt to report figures only for the groups in the main section of this paper and include the (much more numerous) plots for individual subreddits in Supplementary Materials.

## 3.4 Treatment-control matching

For each treatment user, we find a control user that is the most similar to them (Niven et al. 2012; Ali et al. 2018), with similarity being defined in terms of Mahalanobis distance (Stuart 2010), and for computational reasons, we randomly subsample the (initially much larger) pool of potential control users to be ten times the size of the treatment users' pool, this is done separately for each subreddit.

The Mahalanobis distance is calculated using the following user attributes: *account creation date*, *Reddit karma (sum of all up-votes minus all down-votes)*, *total number of submissions*, *total number of comments*, and the *count of posts made in each of the 50 most similar subreddits* (those with the highest ratio of treatment members, as per the list generated when defining the subreddits from which to sample control candidates).

We always cap the matching features at the month prior to the treatment user becoming active on the hate subreddit, that same date cutoff is also applied to all control candidates

[2] The complete hate lexicon: https://github.com/Matheus-Schmitz/Reddit_Hate/tree/main/hate_speech_lexicons.

under consideration for matching. This follows standard causal methods that avoid unintentional correlations between matching and outcome (Ham and Miratrix 2022).

The matching algorithm uses the following procedure: (1) select a treatment user, (2) check in which month that user became active on the hate subreddit, (3) consider that user's features and the control candidates' features on the month prior to the activation month, (4) find the most similar control candidate via Mahalanobis distance matching, (5) store a triplet of (treatment, control, distance), (6) move to the next treatment user.

There is a possibility that more than one treatment user gets matched to the same control candidate, hence once all users were matched, we check for control candidates who had more than one match and keep only the match with the shortest Mahalanobis distance. The remaining de-matched treatment users were put back on the queue to be re-matched. We also then remove those control users who already have a match from the control dataset such that the treatment user will have to find its next best match. This is repeated until each control user has at most one match.

Once complete, the algorithm provides a one-to-one pairing between treatment and control users that contains the most similar pairs possible. For each pair, we center its data, defining as day 0 the day on which the treatment user first posted in the hate community (i.e., when the treatment began), and that same calendar day is also day 0 for the matched control user. In absolute calendar dates, day 0 differs between pairs as different treatment users became active on the hate community on different dates. This makes our analysis robust to disturbances introduced by any single-day event (Pearce 2016).

### 3.5 Interrupted time series

Among the Regression Discontinuity Design family of models, the interrupted time series (ITS) approach was conceived for causal-effect modeling of changes to systems over a defined time span, and has been successfully applied to numerous fields (Ham and Miratrix 2022; Lee and Lemieux 2009; Cattaneo et al. 2020).

#### 3.5.1 Bandwidth optimization

The analysis period in which datapoints will be considered is referred to as the *bandwidth* and is usually symmetric around a date in which the event of interest happened, such that a bandwidth of, e.g., 30 days means an analysis period from -30 to +30 days around the event of interest. In our research, the date for the event of interest (becoming active in a hate subreddit) has been normalized to be day 0, as explained in Sect. 3.4.

To determine the optimal bandwidth, we employ the same quantitative method leveraged in prior research (Imbens and Lemieux 2008; Baicker and Svoronos 2019; Ludwig and Miller 2005), whereby leave-one-out cross-validation is used to select amongst a set of candidate bandwidths. We again follow prior research and use root-mean-squared error (RMSE) (Jacob et al. 2012) as the metric to evaluate and choose amongst the bandwidth candidates.

To ensure stability, the leave-one-out cross-validation analysis is run for 100 rounds. This number was based on choosing one subreddit at random, and verifying the chosen bandwidth when setting the number of rounds to be 5,10,15,...,200, which revealed that roughly past 100 cross-validation rounds results are stable, i.e., we consistently recommend the same bandwidth. Furthermore, since bandwidth optimization should not interfere with the modeling and estimation of interrupted time series parameters, in defining the optimal bandwidth we only consider data points prior to the event of interest, which is to say our datapoints are in the range of $(-n, -1]$ (Turner Simon et al. 2021; Lopez et al. 2018; Ewusie Joycelyne et al. 2017).

We illustrate the process for calculating the RMSE associated with a given bandwidth by using the example case where the candidate bandwidth is ten days, using three cross-validation rounds: (A) fit a linear regression from days $[-11, -2]$; (B) use the fitted model to predict on day $-1$; (C) on two separate (initially empty) arrays, append the datapoints associated with the true value for day $-1$ and the predicted value for day $-1$, respectively; (D) repeat (A–C) using days $[-12, -3]$ to predict for day $-2$; (E) repeat (A–C) using days $[-13, -4]$ to predict for day $-3$, we have now reached our desired number of cross-validation rounds for this given bandwidth; (F) now comparing the array of truth values and the array of predicted values, calculate RMSE, and store it as a RMSE associated with a bandwidth of ten days.

The process described above can then be replicated for the entire set of candidate bandwidths, and once it is complete, we then compare the RMSE of all candidates and choose the one with the lowest value, which means that is the number of days that best models our data before the event of interest occurred (on day 0). In our analysis, we considered bandwidths in the range of 30–365 days in five-day increments. Values below 30 were left out due to providing too few samples for later modeling, and values above 365 days were deemed to be too susceptible to longitudinal factors that could confound with the causal effects being studied. The optimal bandwidth found for each subreddit is shown in Table 1.

The datapoints considered during ITS modeling are limited to posts made by members of the studied subreddit *outside* of the community in question, in order to control for group specific behavior, and because our aim is to measure how hate spreads to the out-group. This only applies to

**Table 2** Interpretation of coefficients obtained from the ITS model

| Parameter | Meaning |
|---|---|
| *const* | Pre-treatment baseline |
| *time* | Pre-treatment trend |
| *expos* | Incremental baseline for the treatment group |
| *inter* | Incremental baseline after treatment |
| *time* × *expos* | Incremental trend for the treatment group |
| *time* × *inter* | Incremental trend after treatment |
| *expos* × *inter* | Incremental baseline on treatment group after treatment |
| *time* × *expos* × *inter* | Incremental trend for the treatment group after treatment |

treatment users, as by definition control users do not have datapoints inside the studied subreddit, since having posts inside the studied subreddit is the definition of what qualified a user as an active member, and all such users are part of the treatment group and not part of the control group.

#### 3.5.2 Model design

ITS is based on ordinary least square (OLS) regression (Turner Simon et al. 2021; Lopez et al. 2018). Its two base variables are *const* and *time*, which measure the baseline hate speech level and the longitudinal trend, respectively. Two dummy variables are added to generate interaction terms for modeling specific subsections of the data. The first of those variables is *exposed*, which indicates whether a datapoint belongs to the control group (exposed = 0) or to the treatment group (exposed = 1). The second dummy variable is *interrupted*, which indicated whether a datapoint belongs to the pre- or post-treatment period. Regardless if treatment or control, all datapoints from day $(-n, -1]$ are pre-treatment and thus have their *interrupted* variable set to 0, and all datapoints from days $[0, n)$ are post-treatment with *interrupted* equals 1.

Next, an ITS model considers all interaction terms between *time*, *exposed*, and *interrupted* (Lee and Lemieux 2009). Table 2 summarizes all ITS coefficients and their interpreted meaning. Through all possible interaction terms, a single ITS model can generate four regression best-fit estimates, accounting for both treatment and control users, on both pre- and post-treatment periods.

#### 3.5.3 Model interpretation

For a given fitted model, it is possible to leverage the obtained parameter coefficients to calculate the relative increase in hate speech when going from pre-treatment to post-treatment (Lopez et al. 2018; Turner Simon et al. 2021). This is done by first obtaining the incremental post-treatment effect size, which for the treatment group is defined by the summation of *inter* and *expos* × *inter*. Next, we obtain the pre-treatment baseline for the treatment group, calculated as *const* + *expos*. We then measure the size of the post-treatment increment in proportion to pre-treatment levels by a simple ratio of *incremental post-treatment effect / pre-treatment baseline*. This gives us the best estimate of the instantaneous change in hate speech that the treatment causes. Note this measure does not use the *time* coefficient and is thus not affected by longitudinal trends. Rather it is a comparison of the best estimate of the hate speech right before and right after becoming active in the hate subreddit (Turner Simon et al. 2021).

When we consider the individual variables in each model, *expos* × *inter* is the most relevant model coefficient, as it captures the instantaneous effect on a user's hate speech as a result of becoming active in the hate community. *inter* would be relevant for a similar reason but is always statistically indistinguishable from zero, as it applies to both treatment and control users, the latter having no statistically significant changes in hate speech post-treatment, and thus leading the coefficient to be zero. If *expos* × *inter* > 0, then the user's hate speech increases as a result of exposure to the hate group, while if *expos* × *inter* ≤ 0, then the exposure would have reduced hate speech.

#### 3.5.4 Sensitivity analysis

To ensure that our results are robust to variations in bandwidth size, we run a sensitivity analysis whereby we fit one ITS model for each possible bandwidth within the range between 30 and 365 days and measure that model's coefficients as well as their p-values (Lopez et al. 2018; Chatterjee and Hadi 2009). This test allows us to observe that for a fixed subreddit dataset the ITS coefficients all converge to consistent values independent of bandwidth size. Yet, sensitivity analysis also highlights one difficulty in the intersection of bandwidth optimization and ITS modeling: models built with a small bandwidth contain fewer datapoints, which increases the variance, leading to higher p-values. P-values then go down in tandem with increase in bandwidth size. As previously explained, our category-based pooling of data is intended to counterbalance this effect (Ewusie Joycelyne et al. 2017; Niven et al. 2012). As a reminder,

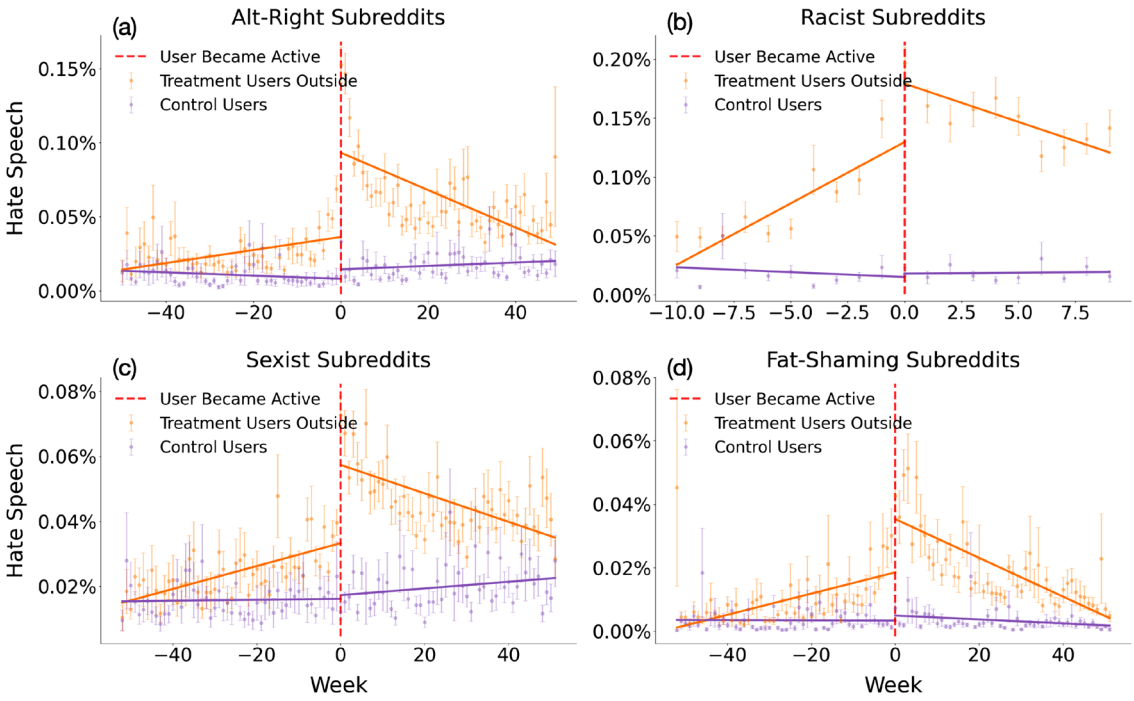


**Fig. 1** Interrupted time series plots for the studied subreddit categories. We plot the rate of hate speech over time (percent of words that are hate words) outside of the subreddit in question for treatment users that join the hate subreddit and control users that never join. The subreddit categories are, in order, **a** alt-right, **b** racist, **c** sexist, and **d** fat-shaming. We consistently observe a pattern of increasing hate speech before users become active in the hate community, followed by a spike when the user becomes active, and then a slow decline in hate speech levels

bandwidth optimization and ITS modeling are performed independently.

## 4 Results

We first explore how hate speech expressed by users changes over time. When we inspect the ITS plots, as shown in Fig. 1, we notice certain trends that replicate across subreddits and categories. First, and unsurprisingly, users who become active on hate subreddits have a higher baseline hate speech even in the pre-treatment period. Figure 1 also reveals that treatment users display an upward trend in their hate speech before becoming active in a hate community, this can be observed by the rising orange line (best fit estimate) for all categories. Despite this trend, ITS's causal-effect modeling at the break-point (interruption, i.e., becoming active) still allows us to infer that there is a causal link between becoming active in the hate community and seeing an immediate increase in hate speech *outside* that community. Yet, this pre-treatment trend is worthy of consideration given that it is consistently observed across all studied subreddits, and we explore it in more detail in the Discussion section.

Using the approach outlined in Sect. 3.5, we estimate the effect that becoming active in a hate community has on users' hate speech outside of the particular subreddit. For each category, we aggregated the time series data of all its subreddits. Figure 2 shows the effect for each hate category, and demonstrates that the effect size is relatively consistent, but is highest for alt-right subreddits. When we disaggregate across categories in Supplementary Fig. S1, we observe similar increases in hate speech, with the largest effect in an alt-right subreddit, r/frenworld. For all subreddits in which the effect size is not statistically significant from zero, we find that the algorithmically-derived bandwidth is especially small (milliondollarextreme = 35 days, CoonTown = 35 days, GreatApes = 80 days, MGTOW = 30 days). For all other subreddits whose algorithmically-derived bandwidth is larger, we observe statistical significance in this plot. More importantly, when we look at Supplementary Fig. S2 (when aggregating across all subreddits in a given category) or Supplementary Fig. S3 (for data disaggregated by subreddit), we can see that $expos \times inter$ reaches statistical significance ($p$-values are below 0.05) on all larger bandwidths. Those plots show that individual model coefficients are consistent across wide variations in bandwidth, indicating that results are not sensitive to one particular bandwidth, for example, looking at r/frenworld on Fig. S3, we observe that once the bandwidth is large enough to be robust to noise, there are no variations in the coefficients anymore, with their associated

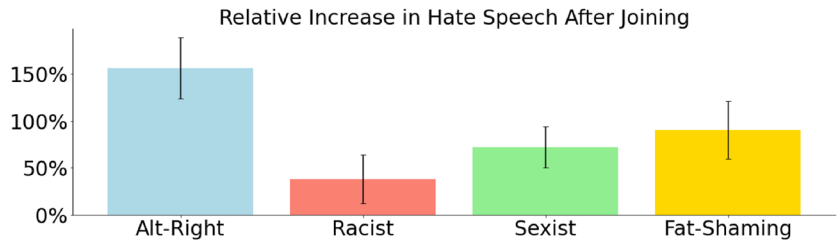


**Fig. 2** Relative increase in the rates of hate speech immediately after users become active in a hate subreddit category, as obtained from the interrupted time series model. All categories observe increases in hate speech, but the magnitude of change varies. Error bars represent 95% confidence intervals

p-values remaining always below 0.05 for the coefficients with a nonzero estimate, whereas the coefficients with a zero estimate continue to be so and to display high p-values.

Wald's *F*-test shows all fitted models are statistically significant ($p$-value $< 10^{-20}$). Our models all have $p$-values $> 0.05$ for all three coefficients (*time*, *inter*, and *time* × *inter*) that apply to all users (treatment and control). We therefore cannot reject the null hypothesis where the real value for those coefficients is zero. Coefficients that apply to all users being statistically near zero lets us know that control users show no statistically significant hate speech changes over time. This result gives us greater confidence that the studied users are not exposed to significant external effects during the analysis period.

Persistent across all studied subreddits is a reduction in hate speech levels after joining hate subreddits, although hate speech levels still remain higher than pre-treatment levels even at the end of the analysis bandwidth. Because Reddit takes an active stance in moderating its platform, Reddit's banning of hate users could be behind this trend, especially when one considers that the users most likely to be struck by a ban are those who use the most hate speech and thus most elevate the treatment groups' hate speech. As these users become banned, the group's hate speech levels trend down.

We assess this hypothesis by investigating the connection between an account's hate speech levels and how long that account continued to show overall Reddit activity after having become active in the hate subreddit. We cannot measure actual bans, but rather the last time in which an account made a submission or comment on the platform. Table 3 shows the hate speech levels for accounts that lasted less versus more than one year after becoming active in the studied hate subreddit. For both groups, we only consider posts after they become active in the hate subreddit, and up to one year from that date. For the group where users remained active on Reddit more than one year after they became active in the hate subreddit, we only consider the posts within the first 365 days after becoming active. We see that, for all considered subreddits, the average hate speech for users who disappear within a year is several times greater than for longer-lasting users. This lends evidence to the "active moderation" hypothesis.

In Fig. 3, we show the percent change in the rate of hate words for users across *all subreddits*, *inside the treatment subreddit*, *outside the treatment subreddit*, *inside banned subreddits*, and *inside non-banned subreddits* for each hate subreddit category. The change in hate speech is most dramatic for the alt-right subreddits, but we see the change appears longest lasting in the sexist subreddits. Supplementary Fig. S4 shows similar results when disaggregated by subreddit. As in the aggregated data, there are dramatic increases in alt-right subreddits like r/frenworld or r/honkler, but long-lasting changes appear in sexist subreddits, such as r/MGTOW.

It becomes clear that there are limitations to considering a single hate subreddit in a vacuum, as before they become active in the studied subreddit, we can observe that treatment users were already using the same hate lingo at higher levels inside other subreddits that were eventually banned. For example, when looking at "Racist Subreddits" (subfigure b), it is clear that the treatment users were already using much higher levels of racist language inside other subreddits that were eventually banned (solid brown line), when compared

**Table 3** Rate of hate speech for hate users who remain for under or over one year

| Subreddit | ≤ 365 days (%) | > 365 days (%) |
|---|---|---|
| r/frenworld | 0.015 | 0.001 |
| r/honkler | 0.444 | 0.124 |
| r/milliondollarextreme | 0.096 | 0.019 |
| r/CoonTown | 0.465 | 0.109 |
| r/GreatApes | 0.421 | 0.125 |
| r/WhiteRights | 0.293 | 0.098 |
| r/Braincels | 0.245 | 0.091 |
| r/Incels | 0.083 | 0.031 |
| r/MGTOW | 0.126 | 0.060 |
| r/fatpeoplehate | 0.078 | 0.014 |

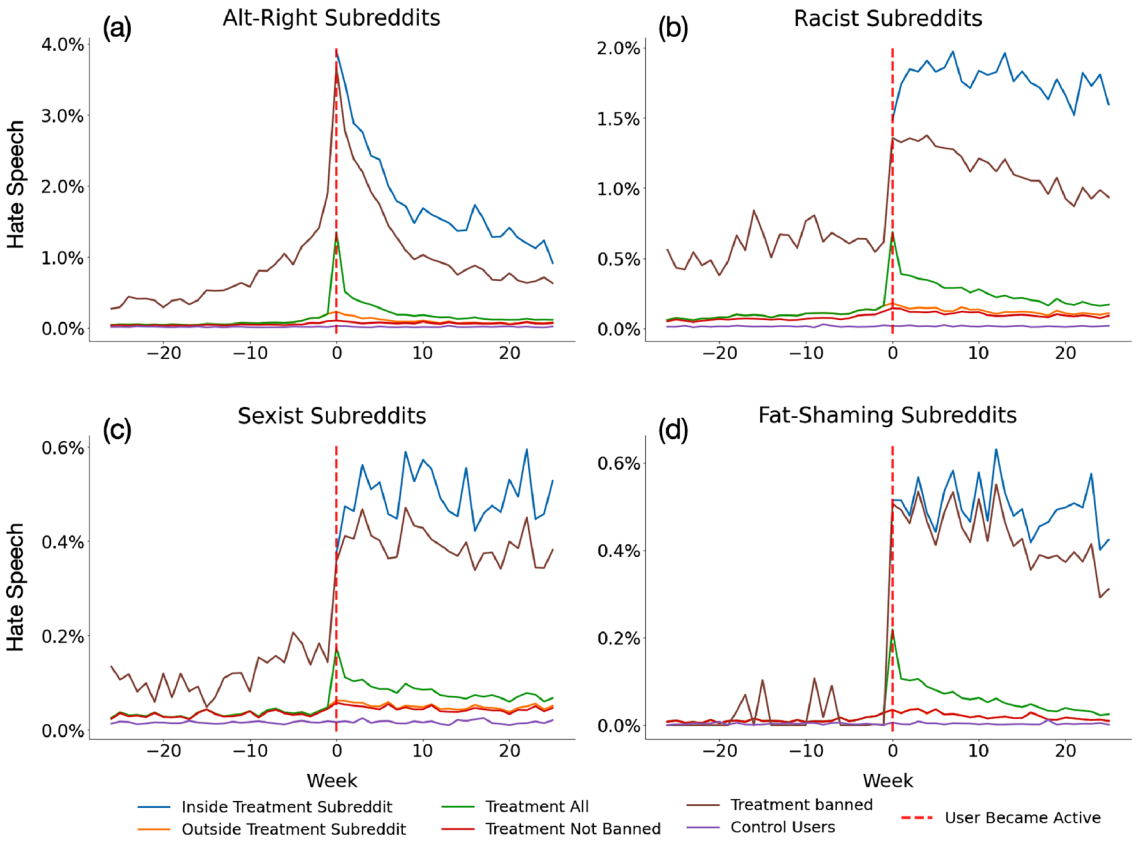


**Fig. 3** Hate speech over time. The percent of all words (comments and submissions) that are hate words for **a** alt-right subreddits, **b** racist subreddits, **c** sexist subreddits, and **d** fat-shaming subreddits. In all plots, we show hate speech for treatment users in the hate subreddit (blue), treatment users outside the hate subreddit (orange), all posts among treatment users (green), all posts by treatment users in non-banned subreddits (red), and all posts by treatment users in subreddits that will eventually be banned as of January 1, 2023 (brown). Control users are in purple. The red dashed line is when hate users first start posting in hate subreddits. (Color figure online)

to baselines such as "control users" (solid purple line) or even the treatment users themselves inside "not banned" subreddits (solid red line). Since all studied subreddits were eventually banned, they are also part of the "banned subreddits" category, and thus it is only natural that category would see a significant spike when the user starts posting in the studied subreddit. A similar logic applies to the "all" category, which contains all subreddits, including the studied one. This gives us the perspective that, for as much as treatment users are nastier than control users in the whole of the platform, when communicating inside hate communities they let loose their worst.

For our analysis, we built a community-specific lexicon, as delineated under Sect. 3.2. A sample of the most popular words from each lexicon is shown in Supplementary Fig. S5, where we consider their usage by hate users both when inside and outside of their communities. Supplementary Figure S5 depicts how treatment users modulate their vocabulary when talking to the in-group and out-group. In more detail, we can see that words most clearly associated with hate, such as the n-word within r/GreatApes and "obeast" within r/fatpeoplehate, double in frequency when users are in the in-group, whereas when communicating to the out-group users shrink their usage of obscure group slang such as "roastie" (used in r/Incels) and "sheboon" (used within r/CoonTown and r/GreatApes). On the whole, we qualitatively observe that hate users favor less intense hate words when in the out-group.

## 5 Discussion and conclusions

We modeled the impact of joining a hate subreddit in terms of users' hate speech in the out-group, unveiling a causal connection between both that is consistent and replicable across communities while also being robust to variation in analysis period, as shown by our sensitivity analysis. Namely, our causal models consistently show that users who become active in a hate community increase their hate speech outside of that community, implying a viral spreading effect on new members. This causal link is shown across ten subreddits and four categories of hate speech.

We began the paper with a tantilizing question: do users adopt extremist beliefs from exposure to hate subreddits? To answer this, we need to determine if users are "...exhibiting intrinsic out-group hostility" and "rejecting egalitarian and democratic values" (Marwick et al. 2022). Our results imply that users become more antagonistic after becoming active in hate groups by using more hate speech in non-hate subreddits. Moreover, the posts users write in hate subreddits tend to demonstrate hatred or even violence towards these out-groups. Both findings lead us to conclude the users exhibit greater out-group hostility. Because we do not know what users think, we cannot directly determine if they are rejecting egalitarian and democratic values more than they did prior to exposure (Marwick et al. 2022), but this is a reasonable if still hypothetical conclusion. Our work therefore provides evidence that users adopt these extremist beliefs simply from exposure to hate subreddits.

Supplementary Figure S5 shows that users exposed to these hate subreddits also tend to use hate words at different frequencies when communicating inside versus outside the hate subreddit. Moreover, in line with prior research (Trujillo et al. 2021), we see an increase in the usage of more cryptic in-group words, such as "foid" (used in the incel communities) or "dindu" (used in racist communities), when communicating with hate group peers, which could be a way of evoking membership through specific language, and thus entice increased responsiveness from other members (Tran and Ostendorf 2016). It is also possible that the relative increase of insider language in the in-group is merely a reflection of the opposite case, where users might in fact be toning down such language in the out-group, for using group-specific language outside might lead to lower engagement, perhaps because the user base at large is not interested in the causes related to the hate words, or merely because they cannot comprehend what is meant by the in-group slangs.

Moreover, when treatment users speak to the out-group, we also observe a reduction in words that qualitatively we perceive as more egregiously offensive (such as using "faggot" instead of the n-word). This could signify that even hate users still feel some of the societal pressures to behave properly, or face harassment. Another explanation is, again, moderation, where it might simply be that users who freely peruse a highly offensive vocabulary are quickly banned from out-group subreddits, and soon cease to be part of our data sample of out-group speech.

There are a few caveats to our research, however. First, despite the matching technique obtaining the best pairing possible, users were not identical in the pre-treatment period, as shown by the differing levels of hate-speech that pre-date treatment. This is caused at least partially from the need not to use the target variable of ITS (hate speech) during matching, to prevent contamination of the results (Ham and Miratrix 2022). As explained in the Methods section, our model is robust to these differences in initial hate speech levels, and can still assign a causal effect to the instantaneous spike in hate speech observed immediately after treatment. Since hate speech exists in a spectrum, yet lexicon based approaches require binary categorizations for matters of practicality, the construction of hate speech lexicons can have a significant impact on the analyses that follow it, given that one might, for example, end up ignoring very large levels of more mild hate speech.

Next, in Fig. 1, we notice an upward trend amongst treatment users before the treatment period. This indicates that before becoming active in a hate subreddit, they already begin to use more hate speech. This might imply that increased hate pushes them into these subreddits, or that users might already have been exposing themselves to the communities material as passive readers, often called "lurkers" on Reddit. As the data that Reddit makes available only tracks user posts, and not what they see, it is difficult to know the degree to which passive reading plays a role. Once again, despite this upwards trend, the causal nature of ITS models at the break-point allows us to attribute the instantaneous spike in hate speech at the moment of becoming active as being precisely due to becoming active. We show the robustness of our findings for individual subreddits in Supplementary Figs. S1 and S6, which display similar data to the figures presented on the main text, but broken down by individual subreddit. Those figures are connected in that the bars shown in Fig. S1 are proportional to the delta between the treatment best fit estimates (orange lines) at day 0 in Fig. S3, for example r/frenworld has the highest relative hate speech increase estimates in Fig. S1, matching it's having the highest delta (discontinuity) between pre- and post-treatment estimates in S3. We specifically show that the ITS regression models (plotted in Supplementary Fig. S6) have consistent discontinuities around the time users post in hate subreddits. These models find a substantial increase in hate speech, shown in Supplementary Fig. S1.

As noted in the initial analyses, there is a mirror case, where the initial post-treatment spike in hate speech subsides over time, although still remaining above pre-treatment levels. Our analysis points to this potentially being caused by accounts with higher hate speech disappearing from the platform earlier than those with lower hate speech, therefore the long-term effect of joining hate groups may be undercounted due to survivorship bias. Although the root cause for that remains uncertain, the most plausible explanation seems to be that we are observing the effects of Reddit's moderation efforts, which are banning the most egregious hate users first, thus bringing the group's average hate speech down, as the plot shows.

In Supplementary Fig. S4, our time series analysis considering various contexts composed by subsets of

subreddits, we observe that those in the banned category has persistently higher levels of hate speech. Since subreddits cannot adopt hate words after their ban, it must be that the observed higher hate speech precedes the ban, although we cannot prove whether those higher hate speech levels cause the ban, or if both are merely correlated due to other underlying factors at play.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s13278-023-01184-8.

**Acknowledgements** Funding for this work is provided through the USC-ISI Exploratory Research Award and through DARPA (Awards # HR0011260595 and # HR001121C0169)

**Author Contributions** All authors designed research; M.S. performed analysis; all authors wrote and reviewed the manuscript.

**Funding** Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

## Declarations

**Conflict of interest** The authors declare no conflicts of interest.



## References

Ali MM, Patrick JN, Sander G (2018) Case-control matching: effects, misconceptions, and recommendations. Eur J Epidemiol 33:5–14

Arguello J, Butler BS, Joyce E, Kraut R, Ling KS, Rosé C, Wang X (2006) Talk to me: foundations for successful individual-group interactions in online communities. In: CHI, pp 959–968

Baicker K, Svoronos T (2019). Testing the Validity of the Single Interrupted Time Series Design. National Bureau of Economic Research, 26080. https://doi.org/10.3386/w26080

Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J (2020) The pushshift reddit dataset. CoRR arXiv:2001.08435

Burke M, Marlow C, Lento T (2009) Feed me: motivating newcomer contribution in social network sites. In: CHI, pp 945–954

Burke M, Settles B (2011) Plugged in to the community: social motivators in online goal-setting groups. In: C &T, pp 1–10

Cattaneo MD, Idrobo N, Titiunik R (2020) A practical introduction to regression discontinuity designs: foundations. In: Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press

Chandrasekharan E, Jhaver S, Bruckman A, Gilbert E (2022) Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit. ACM Transactions on Computer-Human Interaction 29(4):1–26. Association for Computing Machinery (ACM). https://doi.org/10.1145/3490499

Chandrasekharan E, Pavalanathan U, Srinivasan A, Glynn A, Eisenstein J, Gilbert E (2017) You can't stay here: the efficacy of reddit's 2015 ban examined through hate speech. In Proceedings of the ACM on Human-Computer Interaction 1:1–22). Association for Computing Machinery (ACM). https://doi.org/10.1145/3134666

Chatterjee S, Hadi AS (2009) Sensitivity analysis in linear regression. Wiley

Chatzakou D, Kourtellis N, Blackburn J, De Cristofaro E, Stringhini G, Vakali A (2017) Hate is not binary: studying abusive behavior of #gamergate on twitter. In: Proceedings of the 28th ACM conference on hypertext and social media, HT '17, New York. Association for Computing Machinery, pp 65–74

Choi B R, Kraut R E, Fichman M (2008). Matching People And Groups: Recruitment And Selection In Online Games. SIGHCI 2008 Proceedings. 3. https://aisel.aisnet.org/sighci2008/3

Contributors Wikipedia (2022) Wikipedia: please do not bite the newcomers—Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Wikipedia:Please_do_not_bite_the_newcomers. Accessed 8 Oct 2022

Copland S (2020) Reddit quarantined: Can changing platform affordances reduce hateful material online? Internet Policy Rev 9(4):1–26

Danescu-Niculescu-Mizil C, West R, Jurafsky D, Leskovec J, Potts C (2013) No country for old members: user lifecycle and linguistic change in online communities. In: TheWebConf, pp 307–318

Davies C, Ashford J, Espinosa-Anke L, Preece A, Turner L, Whitaker R, Srivatsa M, Felmlee D (2021) Multi-scale user migration on reddit. In: 15th international AAAI conference on web and social media

Eisenstein J, Ahmed A, Xing EP (2011) Sparse additive generative models of text. In: ICML

ElSherief M, Kulkarni V, Nguyen D, Wang WY, Belding E (2018) Hate lingo: a target-based linguistic analysis of hate speech in social media. In: Proceedings of the international AAAI conference on web and social media, vol 12

Ewusie Joycelyne E, Erik B, Charlene S, Joseph B, Lehana T, Straus Sharon E, Hamid Jemila S (2017) Methods, applications, interpretations and challenges of interrupted time series (its) data: protocol for a scoping review. BMJ Open 7(6):e016018

Gallacher J D, Bright J (2021). Hate Contagion: Measuring the spread and trajectory of hate on social media. PsyArXiv. https://doi.org/10.31234/osf.io/b9qhd

Gaudette T, Scrivens R, Davies G, Frank R (2021) Upvoting extremism: collective identity formation and the extreme right on reddit. New Media Soc 23(12):3491–3508

Gerrard Y (2018) Beyond the hashtag: circumventing content moderation on social media. New Media Soc 20(12):4492–4511

Gothard KC (2020) Exploring INCEL language and subreddit activity on reddit. University of Vermont

Halfaker A, Kittur A, Riedl J (2011) Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In: WikiSym, pp 163–172

Ham D W, Miratrix L (2022). Benefits and costs of matching prior to a Difference in Differences analysis when parallel trends does not hold. arXiv:2205.08644. https://doi.org/10.48550/ARXIV.2205.08644

Hickey D, Schmitz M, Fessler D, Smaldino PE, Muric G, Burghardt K (2023) Auditing elon musk's impact on hate speech and bots. In: Proceedings of the international AAAI conference on web and social media, vol 17, pp 1133–1137

Hickey D, Schmitz M, Fessler D, Smaldino P, Muric G, Burghardt K (2023) No love among haters: negative interactions reduce

online hate community engagement. arXiv preprint arXiv:2303.13641
Imbens GW, Lemieux T (2008) Regression discontinuity designs: a guide to practice. J Econometr 142(2):615–635. The regression discontinuity design: Theory and applications
Jacob R, Zhu P, Somers M A, Bloom H (2012). A practical guide to regression discontinuity. MDRC. ERIC Internal Report, No. ED565862: unpublished
Jhaver S, Ghoshal S, Bruckman A, Gilbert E (2018) Online harassment and content moderation: the case of blocklists. TOCHI 25(2):1–33
Johnson Neil F, Rhys L, Johnson RN, Nicholas V, Minzhang Z, Pedro M, Prajwal D, Stefan W (2019) Hidden resilience and adaptive dynamics of the global online hate ecology. Nature 573(7773):261–265
Kraut RE, Resnick P (2012) Building successful online communities: evidence-based social design. MIT Press
Kumar S, Hamilton WL, Leskovec J, Jurafsky D (2018) Community interaction and conflict on the web. In: TheWebConf, pp 933–943
Laeeq KM (2017) Social media engagement: What motivates user participation and consumption on youtube? Comput Hum Behav 66:236–247
Lee DS, Lemieux T (2009) Regression discontinuity designs in economics. Working Paper 14723, National Bureau of Economic Research
Lewis JA, Gee PM, Ho CL, Miller LM (2018) Understanding why older adults with type 2 diabetes join diabetes online communities: semantic network analyses. JMIR Aging 1(1):e10649
Lopez BJ, Soumerai S, Antonio G (2018) A methodological framework for model selection in interrupted time series studies. J Clin Epidemiol 103:82–91
Ludwig J, Miller DL (2005) Does head start improve children's life chances? Evidence from a regression discontinuity design. Working Paper 11702, National Bureau of Economic Research
Marchal N (2020). The Polarizing Potential of Intergroup Affect in Online Political Discussions: Evidence From Reddit R/Politics. In SSRN Electronic Journal. Elsevier BV. https://doi.org/10.2139/ssrn.3671497
Marwick A, Clancy B, Furl K (2022) Far-right online radicalization: a review of the literature. Bull Technol Public Life. https://citap.pubpub.org/pub/jq7l6jny
Massanari A (2017) #gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. New Media Soc 19(3):329–346
Mondal M, Silva LA, Benevenuto F (2017) A measurement study of hate speech in social media. In: Proceedings of the 28th ACM conference on hypertext and social media, HT '17, New York. Association for Computing Machinery, pp 85–94
Natalie B, Jo PC (2012) Pro-anorexia communities and online interaction: Bringing the PRO-ANA body online. Body Soc 18(2):27–57
Newell E, Jurgens D, Saleem H, Vala H, Sassine J, Armstrong C, Ruths D (2016) User migration in online social networks: a case study on reddit during a period of community unrest. In: 10th international AAAI conference on web and social media
Niven DJ, Berthiaume LR, Fick GH, Laupland KB (2012) Matched case-control studies: a review of reported statistical methodology. Clin Epidemiol., 4:99-110. https://doi.org/10.2147/CLEP.S30816; PMID: 22570570; PMCID: PMC3346204.
Pearce N (2016). Analysis of matched case-control studies. In BMJ 352(8046):i969). https://doi.org/10.1136/bmj.i969
Phadke S, Mitra T (2021) Educators, solicitors, flamers, motivators, sympathizers: characterizing roles in online extremist movements. CSCW 5(CSCW2):1–35
Phadke S, Mitra T (2020) Many faced hate: a cross platform study of content framing and information sharing by online hate groups. In: CHI, pp 1–13
Rathje S, Van Bavel J J, van der Linden S (2021). Out-group animosity drives engagement on social media. In Proceedings of the National Academy of Sciences 118(26):e2024292118. https://doi.org/10.1073/pnas.2024292118
Ridings CM, Gefen D (2004) Virtual community attraction: why people hang out online. J Comput-Mediat Commun 10(1):JCMC10110
Rieger D, Kümpel AS, Wich M, Kiening T, Groh G (2021) Assessing the extent and types of hate speech in fringe communities: a case study of alt-right communities on 8chan, 4chan, and reddit. Soc Media Soc 7(4):20563051211052904
Santos T, Burghardt K, Lerman K, Helic D (2020) Can badges foster a more welcoming culture on q &a boards? ICWSM 14(1):969–973
Schmitz M, Muric G, Burghardt K (2022) Quantifying how hateful communities radicalize online users. In: ASONAM 2022. IEEE, pp 139–146
Silva L, Mondal M, Correa D, Benevenuto F, Weber I (2016) Analyzing the targets of hate in online social media. In: 10th international AAAI conference on web and social media
Stockdale Laura A, Coyne Sarah M (2020) Bored and online: reasons for using social media, problematic social networking site use, and behavioral outcomes across the transition from adolescence to emerging adulthood. J Adolesc 79:173–183
Stuart EA (2010) Matching methods for causal inference: a review and a look forward. Stat Sci: Rev J Inst Math Stat 25(1):1–21
Tran T, Ostendorf M (2016) Characterizing the language of online communities and its relation to community reception. In: Proceedings of the 2016 conference on empirical methods in natural language processing, Austin. Association for Computational Linguistics, pp 1030–1035
Trujillo MZ, Rosenblatt SF, Jauregui GD, Moog E, Samson BP, Hébert-Dufresne L, Roth AM (2021) When the echo chamber shatters: examining the use of community-specific language post-subreddit ban. arXiv preprint arXiv:2106.16207
Turner Simon L, Amalia K, Forbes Andrew B, Monica T, Grimshaw Jeremy M, McKenzie Joanne E (2021) Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series. BMC Med Res Methodol 21(1):1–19
Van Der Does T, Galesic M, Dunivin ZO, Smaldino PE (2022) Strategic identity signaling in heterogeneous networks. Proc Natl Acad Sci 119(10):e2117898119
Yao Z, Yang D, Levine JM, Low CA, Smith T, Zhu H, Kraut RE (2021) Join, stay or go? A closer look at members' life cycles in online health communities. CHI 5(CSCW1):1–22
Zannettou S, Bradlyn B, De Cristofaro E, Kwak H, Sirivianos M, Stringini G, Blackburn J (2018) What is Gab: a bastion of free speech or an alt-right echo chamber. In: WWW, pp 1007–1014
Zannettou S, Caulfield T, De Cristofaro E, Kourtelris N, Leontiadis I, Sirivianos M, Stringhini G, Blackburn J (2017) The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In: Proceedings of the 2017 internet measurement conference, IMC '17, New York. Association for Computing Machinery, pp 405–417
Zhang X, Zhu F (2006) Intrinsic motivation of open content contributors: the case of Wikipedia. In: Workshop on information systems and economics, vol 10. Citeseer