You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech

ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology
UMASHANTHI PAVALANATHAN, Georgia Institute of Technology
ANIRUDH SRINIVASAN, Georgia Institute of Technology
ADAM GLYNN, Emory University
JACOB EISENSTEIN, Georgia Institute of Technology
ERIC GILBERT, University of Michigan

In 2015, Reddit closed several subreddits—foremost among them r/fatpeoplehate and r/CoonTown—due to violations of Reddit's anti-harassment policy. However, the effectiveness of banning as a moderation approach remains unclear: banning might diminish hateful behavior, or it may relocate such behavior to different parts of the site. We study the ban of r/fatpeoplehate and r/CoonTown in terms of its effect on both participating users and affected subreddits. Working from over 100M Reddit posts and comments, we generate hate speech lexicons to examine variations in hate speech usage via causal inference methods. We find that the *ban worked for Reddit.* More accounts than expected discontinued using the site; those that stayed drastically decreased their hate speech usage—by at least 80%. Though many subreddits saw an influx of r/fatpeoplehate and r/CoonTown "migrants," those subreddits saw no significant changes in hate speech usage. In other words, other subreddits did not inherit the problem. We conclude by reflecting on the apparent success of the ban, discussing implications for online moderation, Reddit and internet communities more broadly.

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: online communities; hate speech; moderation; banning; causal inference.

ACM Reference Format:
Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 31 (November 2017), 22 pages.
https://doi.org/10.1145/3134666

1 INTRODUCTION

Reddit is organized into over one million[1] user-created and user-moderated communities known as *subreddits*. Alongside mainstream subreddits for discussing scientific discoveries (r/science) and affordable fashion choices (r/frugalmalefashion), Reddit has also seen an increase in "toxic" subreddits—subreddits that exist to target hate speech at certain groups [20]. In response, the site introduced a new anti-harassment policy in 2015 [35]. On June 10, 2015, Reddit took action, announcing that it would ban several subreddits under the new policy [17]. Among them were two

[1] http://redditmetrics.com/history

2573-0142/2017/11-ART31
https://doi.org/10.1145/3134666

notorious subreddits: r/fatpeoplehate and r/CoonTown [4]. In this paper, we study the effectiveness of this ban. (To describe these subreddits, we include examples of hateful content, which readers may find upsetting. However, they are necessary to understand Reddit's response, and to ground our research.)

r/fatpeoplehate was a fat-shaming subreddit devoted to posting pictures of overweight people for ridicule [36]. It was one of the most prominent removals from Reddit, with over 150,000 subscribers at the time of the ban.[2] According to the subreddit's own rules, r/fatpeoplehate users were prohibited from any "fat sympathy" [14]. Provided as an example, the following highly-upvoted r/fatpeoplehate comment was typical on the subreddit:

> "You fucking fatass, you made the decision to be a fat fuck after you decided to stuff your fat fucking face instead of acting like a normal human being."

r/CoonTown was a racist subreddit dedicated to violent hate speech against African Americans. It contained "a buffet of crude jokes and racial slurs, complaints about the liberal media, links to news stories that highlight black-on-white crime or Confederate pride, and discussions of black people appropriating white culture" [28]. Their banner featured a cartoon of a black man hanging, with a Klansman in the background [20]. It had over 20,000 subscribers at the time of banning.[3] The following is a representative, highly-upvoted comment from the subreddit:

> "It would be so much easier if this [n-word] was taken outside and shot. Then rasslle up his eight or nine [kids] and shoot them so we can terminate that line of genes."

1.1 The Effectiveness of the Ban

Any site that allows user contributions struggles with offensive content it would rather not host—for legal, ethical and public relations reasons. On one hand, many internet platforms subscribe to generous free speech principles. On the other, many platforms would also rather not host and financially support—through server and bandwidth fees—groups such as r/fatpeoplehate and r/CoonTown. Apart from the philosophical quandaries surrounding banning, Reddit's decision to ban these deviant hate groups provides us with a unique opportunity to study the *efficacy* of banning as a moderation approach. It is a quasi-experiment through which we can examine the effectiveness of banning as a strategy.

The subject of little empirical study, banning deviant groups from an online community might diminish the behavior, or it may just spread it to other parts of the community. For instance, the well-known "take it outside" design guideline [21] would argue that the existence of spaces such as r/fatpeoplehate and r/CoonTown might help relegate hateful behavior to those parts of Reddit. This paper examines Reddit's decision to take those spaces away: we investigate the longitudinal, causal effects of Reddit's decision to ban the deviant hate groups r/fatpeoplehate and r/CoonTown. By analyzing temporal data via causal inference methods, we aim to causally attribute subsequent changes to the ban.

1.2 Research Questions & Findings

We analyze the effects of the ban at two levels: the *user level* and the *community level.*

> **RQ1: What effect did Reddit's ban have on the contributors to banned subreddits?**
> RQ1a: How were their activity levels affected?
> RQ1b: How did their hate speech usage change, if at all?

[2] http://redditmetrics.com/r/fatpeoplehate

[3] http://redditmetrics.com/r/CoonTown

RQ2: What effect did the ban have on subreddits that saw an influx of banned subreddit users?
RQ2a: To which subreddits did the contributors to banned subreddits migrate after the ban?
RQ2b: How did hate speech usage by migrants change in these subreddits, if at all?
RQ2c: How did hate speech usage by preexisting users change in these subreddits, if at all?

RQ1 aims to understand the effects on users directly involved; whereas RQ2 investigates the second-order effects of closing these subreddits (i.e., "Did Reddit 'spread the infection'?"). We answer our research questions using observational data from Reddit, through temporal analysis of Reddit timelines (all comments and submissions made in 2015). Working from over 100M Reddit posts and comments, we generate hate speech lexicons to examine variations in hate speech usage via causal inference methods.

Within the frame RQ1 and RQ2 provide, *we find that the ban worked for Reddit.* Many more accounts than expected discontinued their use of the site; and, among those that stayed active, there was a drastic decrease (of at least 80%) in their hate speech use. Though many subreddits saw an influx of r/fatpeoplehate and r/CoonTown "migrants," those subreddits saw no significant changes in hate speech use. In other words, other subreddits did not inherit the problem. We conclude by reflecting on the apparent success of the ban. We note that while the ban may have worked *for Reddit*, from a macro-perspective, it may have also relocated the behavior onto other sites.

2 BACKGROUND

Next, we survey research in three topics related to the work presented in this paper: *online moderation*, *hate speech*, and related *migration and matching studies.*

2.1 Online Moderation

There are a variety of different approaches to regulate behavior in online communities. In a comprehensive meta-analysis, Kiesler et al. present ways to limit the damage that bad behavior causes when it occurs, and to limit the amount of bad behavior that a bad actor can perform [21]. Prior work on moderation has primarily focused on the observable effects of social feedback mechanisms (e.g., [10, 24, 27]). A wide range of hateful behavior can be destructive to online communities; however, such behavior is also celebrated in some communities including 4chan [3] and Something Awful Forums [31]. While some kinds of moderation can be effective [41], moderation can also make things worse [7, 9]. Therefore it is unclear if banning will lead to positive changes within a community—the central question of the present work.

Research on automatic approaches to moderating online antisocial behavior has shown that textual cyberbullying [15, 49] and undesirable posting [6, 8, 10, 43] can be identified based on topic models, presence of insults and user behavior. However, the literature in online moderation lacks empirical studies about the effectiveness of various abusive content moderation strategies. This is largely due to the fact that when a site employs a moderation approach that removes content from the internet, it is therefore no longer visible to researchers. The ban of r/fatpeoplehate and r/CoonTown, along with the data used in this paper, gives us a quasi-experiment through which we can empirically study the long-term effects of abusive content moderation in online communities.

2.2 Hate Speech and Online Abusive Language

Although the term *hate speech* is used frequently, there is no universally accepted definition of the term. The Manual on Hate Speech of the European Court of Human Right provides the following operational definition:

> "The term *hate speech* shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin" [48].

Our use of the term. This definition of hate speech is not limited to speech that incites violence—it includes all speech that incites hatred on the basis on various personal characteristics and group membership. An open question is whether this definition of hate speech pertains to body characteristics such as "fatness;" the definition presents a list of such characteristics (minorities, migrants, etc), but it does not stipulate that this list is exclusive. It is clearly the case that racial, ethnic, and homophobic hate speech have well-documented connections to violence and discrimination in the real world. Nonetheless, in this context, we feel that the term "hate speech" is a more accurate description of the content of r/fatpeoplehate than milder alternatives such as "offensive speech" or "abusive language." Speech or writing may be "offensive" to some readers for any number of reason—such as the presence of swear words. Similarly "abusive language" might focus on idiosyncratic personal characteristics that are unrelated to larger social group dynamics. In contrast, r/fatpeoplehate focuses exclusively on denigrating fat people *as a group*.

Prior research on the automated detection of hate speech obtained annotations using slightly different definitions such as "hateful or antagonistic responses with a focus on race, ethnicity, and religion" [5, 46], "messages with abusive or hostile words and phrases" [34, 50], and classifications such as racist/non-racist [23]. Researchers annotated tweets containing hate speech using critical race theory [47]. But we could not use the annotated tweets from this work because most of them were subsequently removed by Twitter.

We take a different, usage-based approach to identify hate speech. First, we automatically extract terms which are unique to the two subreddits that were banned due to hate speech and harassment. The resulting term list includes a number of words that indicate hate speech, as well as some other terms that appear to be specific to the Reddit context. We then qualitatively filter these lists, obtaining a high precision hate lexicon. These lexicons are publicly available to the community as a resource.

2.3 Migration and Matching Studies

The availability of large-scale observational data from social platforms has lead to an increased interest in studying changes in user and community behavior due to external events. Recent work along these lines has for example focused on the changes in social network structure (e.g., [37]) as well as the content and user population (e.g., [29]). Closely related to the external event we consider in this paper, Newell et al. analyzed migration patterns of Reddit users to alternative platforms such as Voat, Snapzu and Empeopled during community unrest in 2015—a finding we revisit at the end of this paper [29] .

While social media data is a rich resource to study naturally occurring social phenomena, there are some challenges when using observational data. Because observational data is not collected under controlled settings, there are potential confounds. Recent studies using large scale observational data have attempted to reduce the effects of such confounds by using techniques from the causal inference literature, such as matching [39] and stratification [19]. Matching techniques have been used to create treatment and control groups of users in studies focusing on social phenomena such as online antisocial behavior [10], mental health [16], dietary choices [12], and weight loss [11]. In our work, we use Mahalanobis Distance Matching [40], to construct a set of control users who have similar characteristics as the treatment users. This helps us make causal inferences about

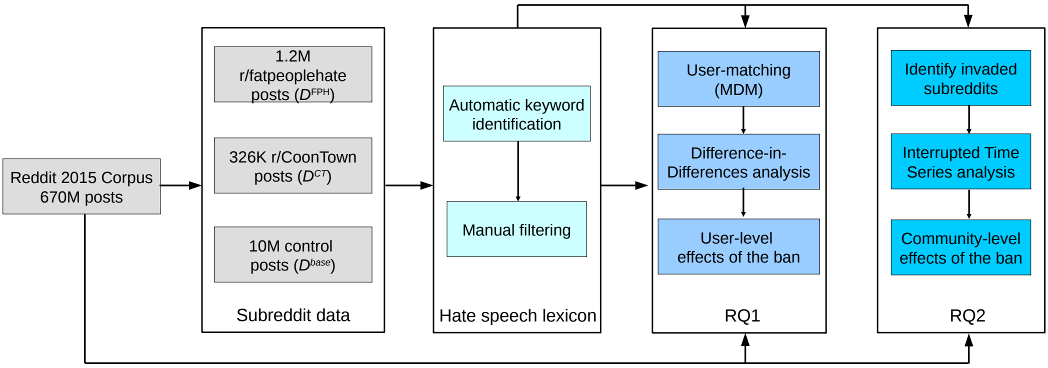
Reddit 2015 Corpus 670M posts
1.2M r/fatpeoplehate posts (D^FPH)
326K r/CoonTown posts (D^CT)
10M control posts (D^base)
Subreddit data
Automatic keyword identification
Manual filtering
Hate speech lexicon
User-matching (MDM)
Difference-in-Differences analysis
User-level effects of the ban
RQ1
Identify invaded subreddits
Interrupted Time Series analysis
Community-level effects of the ban
RQ2

Fig. 1. Flowchart depicting the different components of our research approach.

Table 1. Overview statistics of banned subreddit data.

Subreddit	Posts	Users	Ban date
r/CoonTown	326,776	3,347	Aug 1, 2015
r/fatpeoplehate	1,213,034	22,211	Jun 10, 2015

the effects of ban, while controlling for confound effects. For some of our research questions, we additionally employ a *Difference-in-Differences* strategy to remove the influence of time-invariant omitted confounds [1].

3 DATASETS AND METHODS

Next, we transition to our dataset construction, and then describe the procedure to generate lexicons of hate words. We use the hate lexicons to perform our language analysis.

3.1 Reddit 2015 Corpus

We construct a dataset that includes all posting activities on Reddit in 2015, using publicly available data containing all submissions and comments data extracted from Reddit.[4] We use the textual content obtained from nearly 670M submissions and comments posted between January and December 2015. In the remainder of this paper, we refer to submissions and comments together as "posts." We obtain user and subreddit timelines from this corpus for subsequent analysis.

3.2 Banned Subreddit Data

Using the Reddit 2015 Corpus, we collect all posts made in 2015 from two banned subreddits considered in this paper: r/fatpeoplehate and r/CoonTown. We refer to the datasets of posts from these forums as $\mathcal{D}^{\text{FPH}}$ and $\mathcal{D}^{\text{CT}}$. Descriptive statistics of the data from these two subreddits are provided in Table 1. By extracting the text contained in posts from these subreddits, we generate text corpora for building lexicons of hate speech.

[4] Comments and submissions were queried through the Reddit API and stored by pushshift.io. More details on the publicly available Reddit dataset can be found at https://pushshift.io/using-bigquery-with-reddit-data. While debate exists around the applicability of Reddit's terms to this dataset, we believe that the dataset and its collection methods comply with Reddit's API terms. See https://www.reddit.com/wiki/api-terms for futher information.

3.3 Identifying Hate Speech

A methodological challenge for this research is to determine the impact of Reddit's actions on the prevalence of hate speech throughout the platform. Hate speech and harassment are contentious topics, lacking clear definitions. As discussed above, the European Court of Human Rights notes that "no universally accepted definition of the term 'hate speech' exists" [48]. They adopt a definition including "comments which are necessarily directed against a person or particular group of people", focusing on race, religion, "aggressive nationalism and ethnocentrism", and homophobic speech [48].

This definition provides a useful starting point, but it is difficult to operationalize at scale. We therefore take a usage-based approach: given that Reddit has banned the r/fatpeoplehate and r/CoonTown forums, we focus on textual content that is distinctively characteristic of these forums. Using an automated keyword identification technique, we build lexicons of keywords for r/fatpeoplehate and r/CoonTown, which makes it possible to track whether the words in these lexicons become more common in other forums after the ban. Next, we manually inspect the automatically generated lexicons, and identify a subset of terms that are especially oriented towards hate speech. These manually refined lexicons are sparser, but offer higher precision.

3.3.1 Automatic Keyword Identification: SAGE Analysis. To automatically identify keywords that characterize the forums r/CoonTown and r/fatpeoplehate, we use datasets of posts from these forums, $\mathcal{D}^{\mathrm{CT}}$ and $\mathcal{D}^{\mathrm{FPH}}$. As a baseline comparison, we also build a dataset from a random sample of posts throughout Reddit, $\mathcal{D}^{\mathrm{base}}$. For keyword identification, we limit consideration to content posted before the date of the earliest ban: June 10, 2015. Our goal is to identify terms whose frequencies are especially large in $\mathcal{D}^{\mathrm{CT}}$ and $\mathcal{D}^{\mathrm{FPH}}$, in comparison to $\mathcal{D}^{\mathrm{base}}$.

Due to the long-tail nature of word frequencies [51], straightforward comparisons often give unsatisfactory results. The difference in word frequencies between two groups is usually dominated by stopwords: a 1% difference in the frequency of 'and' or 'the' will be larger than the overall frequency of most terms in the vocabulary. The ratio of word frequencies—equivalent to the difference in log frequencies and to pointwise mutual information—has the converse problem: without carefully tuned smoothing, the resulting keywords will include only the lowest frequency terms, suffering from high variance. The Sparse Additive Generative Model (SAGE) offers a middle ground, selecting keywords by comparing the parameters of two logistically-parametrized multinomial models, using a self-tuned regularization parameter to control the tradeoff between frequent and rare terms [18]. SAGE has been used successfully for the analysis of many types of language differences, including age and gender [33], politics [42], and online discussions of various illegal drugs [32].

We use the Python SAGE implementation[5] to perform two comparisons: $\mathcal{D}^{\mathrm{FPH}}$ versus $\mathcal{D}^{\mathrm{base}}$, and $\mathcal{D}^{\mathrm{CT}}$ versus $\mathcal{D}^{\mathrm{base}}$. In each comparison, we consider the 100 terms with the highest SAGE coefficients.[6] In both cases, the subreddit names themselves are ranked at or near the top, which provides face validity for the keyword identification method. (In the next subsection, we manually remove such self-referential terms.) In r/CoonTown, the remaining terms include a number of words that are either racial slurs, or are terms that frequently play a role in racist argumentation (e.g., 'negro', 'IQ', 'hispanics', 'apes'). In r/fatpeoplehate, the top terms include slurs (e.g., 'fatties', 'hams'), terms that frequently play a role in fat shaming (e.g., 'BMI', 'cellulite'), and a cluster of terms that relate, self-referentially, to the practice of posting hateful content (e.g., 'shitlording', 'shitlady').

3.3.2 Manual Filtering. As noted above, several of the terms generated by SAGE are only peripherally related to hate speech. These include references to the names of the subreddits (e.g.,

[5] https://github.com/jacobeisenstein/SAGE/tree/master/py-sage

[6] Post-hoc robustness checks show that the results are broadly similar for other numbers of terms.

'fph'), references to the act of posting hateful content (e.g., 'shitlording'), and terms that are often employed in racist or fat-shaming, but are frequently used in other ways in the broader context of Reddit (e.g., 'IQ', 'welfare', 'cellulite'). To remove these terms, the authors manually annotated each element of the top-100 word lists. Annotations were based on usages in context: given ten randomly-sampled usages from Reddit, the annotators attempted to determine whether the term was most frequently used in hate speech, using the definition from the European Court of Human Rights mentioned above.

Each term was annotated separately by two independent raters; ties were then broken by a third rater. After labeling was complete, we computed the inter-rater reliability using Cohen's κ, which indicated high inter-rater agreement: 0.875 for r/fatpeoplehate and 0.893 for r/CoonTown hate words. We obtained a total of 23 words with a score of 1.0 (definitely hate) for r/CoonTown and a total of 18 words with a score of 1.0 (definitely hate) for r/fatpeoplehate. The full term lists and annotations are available online.[7]

The manually-filtered keywords lists offer a higher precision estimate of the rate of hate speech, in comparison with the automatically-generated SAGE terms. However, because the manually filtered lists are relatively sparse, estimates of the frequency of hate speech from these lists suffer from high variance. We therefore report comparisons using both the automatically-generated and manually-filtered word lists. Manual filtering removes false positives: terms that are frequently used in hate speech forums, but are not intrinsically hate speech. A more difficult challenge is to identify false negatives, which are terms that convey hate speech despite not being detected by SAGE as high-frequency terms in r/fatpeoplehate and r/CoonTown. Furthermore, abusive language is far more complex than the use of specific words or phrases; identifying such content requires complex linguistic reasoning to determine the author's intent and the message's likely interpretation [30]. This is a long-term challenge for natural language processing, and our keyword-based approach represents only a first step.

4 RQ1: USER-LEVEL EFFECTS OF THE BAN

Next, we explore the user-level effects of the ban, through the following research questions:

> **RQ1: What effect did Reddit's ban have on the contributors to banned subreddits?**
> RQ1a: How were their activity levels affected?
> RQ1b: How did their hate speech usage change, if at all?

4.1 Overarching User-matching Strategy

The causal inference question is whether the banning of a subreddit causes a decrease in posting volume and hate speech usage by users from the subreddit. Ideally, from a study design perspective, Reddit would have randomly chosen the subreddits to ban from a list of candidate subreddits. However, the r/fatpeoplehate and r/CoonTown subreddits were not randomly chosen (to our knowledge), so we employ a number of techniques to approximate the results we would have seen if they had been randomly chosen. These techniques include: matching the treatment subreddits (r/fatpeoplehate and r/CoonTown) to control subreddits that could potentially have been banned, matching the treatment subreddit users to control subreddit users with similar posting behavior, and using a difference-in-differences procedure to compare the pre- and post-differences between the treatment and control groups.

[7] The complete term lists, which contain offensive content, can be found at https://tinyurl.com/hatewords.

Table 2. Examples of the subreddits that were used for generating the pool of control group candidates. As the titles suggest, these subreddits are similar to the banned subreddits.

Subreddit	Co-posting by
FatAcceptanceMovement	FPH
TPWISAFUCKINGBITCH	FPH
ShitlordLife	FPH
HamplanetHateMail	FPH
Fat[N-word]Hate	CT
WhitesWinFights	CT
Watch[N-word]Die	CT
GasTheK*kes	CT

4.2 Treatment Users: Members of Banned Subreddits

Using the subreddit timelines from $\mathcal{D}^{\text{FPH}}$ and $\mathcal{D}^{\text{CT}}$, we mine the user handles of all users who posted in r/fatpeoplehate and r/CoonTown. In order to account for chance posts made by random users, we only consider users who had at least five posts in these subreddits. These users constitute the treatment group, and we refer to them as *treatment FPH* and *treatment CT* users. The treatment applied to these users is that the subreddit they used to post on is banned by Reddit.

4.3 Control Group Candidates: Co-posting with Treatment

At a high level, we generate control users by identifying people who post in other subreddits that are also frequented by treatment users. We first compile all Reddit posts made in unbanned subreddits between January 2015 and June 2015 (pre-ban). In particular, we collect posts from subreddits other than r/fatpeoplehate and r/CoonTown, which were "highly likely" to be banned. An intuitive approach to identify such subreddits would be based on hate speech usage within those subreddits. But we face a circular issue: our operationalized hate speech depends on the lexicons built specifically using posts from r/fatpeoplehate and r/CoonTown. This results in high variance, sparsity and (most problematically) selection on the dependent variable issues. As a result, we instead use the co-posting behavior of treatment FPH and treatment CT users as a proxy for the likelihood of being banned by Reddit.

We compile a list of all subreddits where treatment users post pre-ban, and pick the top 200 subreddits based on the percentage of treatment users posting in these subreddits. Examples of the subreddits that were picked are shown in Table 2 for reference.[8]

Next we obtain the timelines of these subreddits, and extract the user handles of all users who had posted in these subreddits before the ban (excluding treatment user handles). We use this set of control group candidates to compile our set of control users, who are similar to treatment users from r/fatpeoplehate and r/CoonTown. Using this method, we construct a pool of 340,093 users for r/CoonTown and 270,435 users for r/fatpeoplehate, who serve as candidates for the control group, during the user matching step discussed next.

4.4 User-matching: Mahalanobis Distance Matching

Analogous to a classical experiment, our approach is to obtain a control user for every treated user, with similar user characteristics. To obtain a less biased estimate of whether the *treatment* of participating in a subreddit that gets banned has an effect, we apply Mahalanobis Distance Matching (MDM) [40]. We use MDM to obtain a set of control users from the Control group

[8] For subsequent analysis, we do not consider subreddits from this list that were found to be banned (or non-existent) at the time of our data collection.

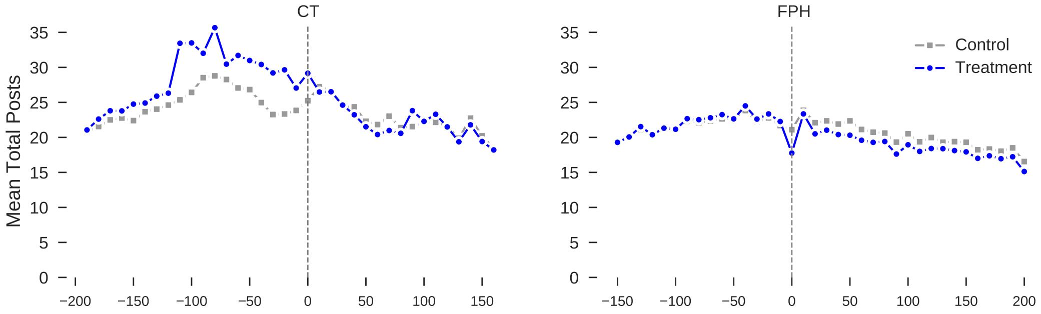

Fig. 2. Variations in users' posting activity on Reddit with respect to the ban. We computed the average number of posts on Reddit by treatment and control users, using time-windows of 10 days spanning 2015, before and after the ban.

candidates mentioned earlier. In MDM, we measure the distance of the users based on three user characteristics (converted to log scale):

account age: days since user account was created
karma: sum of scores on all comments made by the user
total posts: total number of posts made pre-ban in 2015[9]

Next, we match each *treated* user with the nearest *non-treated* (control) user. Finally, we perform Mann-Whitney U tests [13] to measure the goodness of fit, and ensure that we obtain a valid match. We obtained $U > 5,599,800$, p-value > 0.87 for all three covariates of CT users, and $U > 246,620,000$, p-value > 0.82 for all three covariates of FPH users. This indicates strong evidence that there is no significant differences between the three user characteristics of treatment and matched control users. In further analysis, we examine whether the act of receiving the treatment affects posting behavior by analyzing the timelines of all users in both treatment and control groups.

4.5 Temporal Analysis of User Timelines

We examine the user-centric effects of Reddit's ban through temporal analysis of users' posting volumes and hate speech usage. We begin by splitting the user timelines into two time periods with T_{ban} (i.e., time of subreddit ban), shown in Table 1, as origin: *pre-ban* and *post-ban*, using the time of creation (*created_utc*) of each post. Then, we bin posts into time-windows of 10 days. We perform our analysis using these binned posts from user timelines.

4.6 User Activity: RQ1a

First, we identify user accounts that were deleted (at the time of data collection) using the Reddit API. Separately, we compare variations in the posting volume of users around the time of the ban. Given a user handle, we compute the number of posts by the user across the two time periods: *pre-ban* and *post-ban*. In particular, we compute the number of posts by a user, aggregated using 10-day windows. This gives us an estimate of the impact of Reddit's ban on the posting behavior of treatment users.

[9] *total posts* accounts for posts made by a user account in all of Reddit, and is not restricted to any particular subreddit(s).

Table 3. Percentage of users from each group who became inactive post-ban. The percentage of user accounts that were found to be deleted are also shown. The differences in deletions and inactivity between treatment and control users were found to be highly significant using proportion tests and permutation tests.

Group	Total	Inactive	Deleted
Treatment FPH	22,211	21.15%	12.00%
Control FPH	22,211	10.49%	11.18%
Treatment CT	3,347	19.33%	21.30%
Control CT	3,347	15.69%	13.39%

4.7 Results: RQ1a

As shown in Table 3, a sizable number of users from the banned subreddits became inactive, no longer posting on Reddit after the ban(s). The differences in account deletions and inactivity between treatment and control groups, for both FPH and CT users, were found to be significant using proportion tests. In particular, we used 2-sample tests for equality of proportions without continuity correction. Through these proportion tests, we obtained $\chi^2 \geq 1528.9$, p-value $< 2.2e-16$ for all 4 proportions: treatment FPH vs control FPH deletions, treatment FPH vs control FPH inactivity, treatment CT vs control CT deletions, and treatment CT vs control CT inactivity. The number of treatment FPH users who became inactive post-ban were twice the number of control FPH users who became inactive. A similar trend is observed when comparing account inactivity among CT users. The number of treatment CT users who deleted their accounts post-ban was almost twice the number of control CT users who deleted their accounts. Additionally, we performed permutation tests as a robustness check. The results from the permutation tests indicated strong evidence that the effects were caused by the ban (p-value≈ 0.001 for inactivity among CT users, and p-value≈ 0.001 for deletions among FPH users). A detailed description of the procedure and interpretation of the corresponding p-values can be found in *Appendix A*.

As visually apparent in Figure 2, however, there were no drastic differences between the preban and postban posting volumes of active users from both groups—for those users who remained on Reddit. By performing permutation tests, we found no significant evidence that the observed decrease in posting volumes of treatment (both FPH and CT) was caused by the ban (p-value≈ 0.637 for CT users, and p-value≈ 0.897 for FPH users). In other words, the decrease in treatment posting activity in Figure 2 is closely mirrored by the control, reflecting a deeper, underlying pattern unrelated to the ban.

4.8 Hate Speech Analysis: RQ1b

Next, we examine the hate speech usage of treatment and control users. We use both automatically-generated and manually-filtered hate words, as the manually-filtered word lists are relatively sparse. In particular, we calculate the frequency of occurrence of words from the hate lexicon, which we normalize by the total number of words used in posts. We confirm that any observed changes in hate speech usage were caused by the ban through causal inference techniques.

4.9 Results: RQ1b

The amount of hate speech used by treatment users decreased dramatically following the ban. We analyzed over 2.5 million posts by treatment CT and control CT users, and over 13 million posts by treatment FPH and control FPH users. The temporal variations in hate speech usage of treatment and control users are shown in the blue lines of Figure 3. They depict decreases of at least 80% in

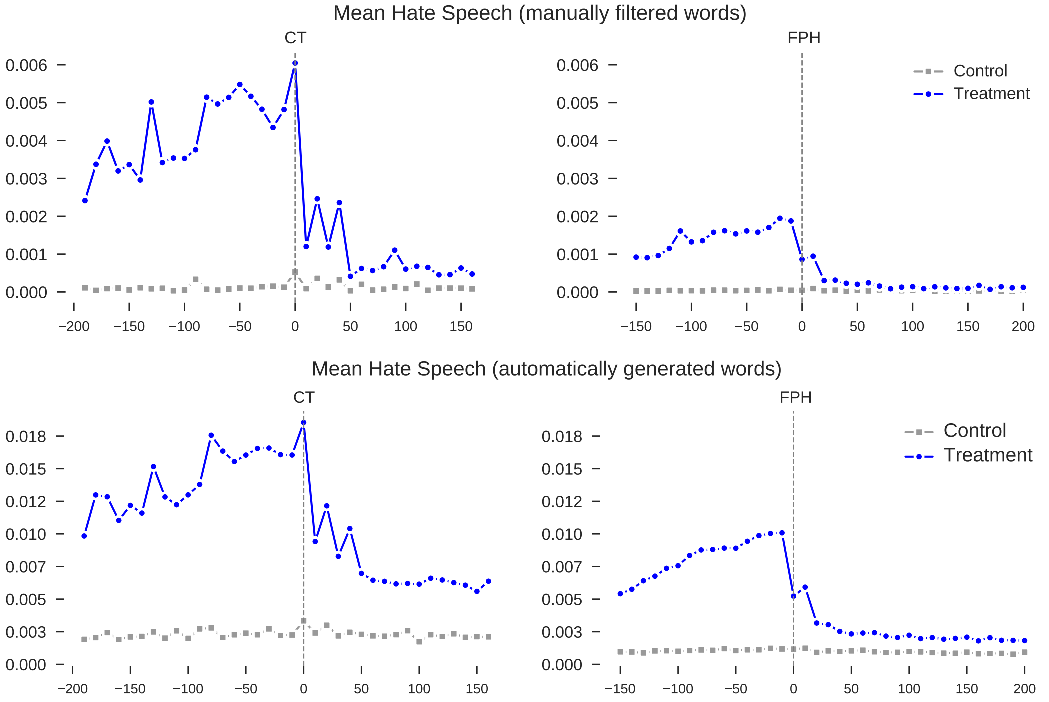

Fig. 3. Variations in users' hate speech usage on Reddit. We compute mean hate speech usage on Reddit by treatment and control users across all of 2015, using time-windows of 10 days, before and after the respective bans. We calculate hate speech usage as the sum of individual frequencies of each term in the hate lexicon and normalize it per post.

treatment groups. However, in order to confirm that these decreases were due to the ban and not some underlying, site-wide decrease in hate-speech behavior, we employ a difference-in-differences analysis as a robustness check.[10]

4.9.1 Difference-in-Difference Analysis for Robustness Check. Matching allows us to control for some measured confounders, but to address the possibility of time-invariant unmeasured confounding, we performed a *Difference-in-Difference* (DiD) analysis [1]. Specifically, DiD works by comparing the differential over-time effect of a treatment on a *treatment group* versus the over-time effect on the *control group*. Because the over-time effect on the control group should be zero (in theory), any estimated effect for the control group will represent bias that should be subtracted from the estimated effect for the treatment group. Informally, the logic of this analysis can be seen in Figure 3, where the large post-ban drop in the blue (treatment) lines is not matched by a post-ban drop in the gray (control) lines.

In order to formally conduct a DiD analysis, we fit the following linear regression model:

[10] The exact causal question is complicated by the fact that the ban likely caused some individuals to become inactive or delete their accounts. The RQ1b analysis implicitly assumes two things: that those treatment users that kept an active account would have kept an active account had their subreddit not been banned, and that those control users that kept an active account would have kept an active account had their subreddit been banned.

Table 4. Results of RQ1 analysis. Hate speech usage measured using both manually filtered and automatically generated hate words are reported. diff_{pre} and $\text{diff}_{\text{post}}$ refer to the difference in labeled hate speech usage between treatment and control users computed before and after the ban. Δ refers to the relative change in these differences following the ban. DiD coef is the coefficient of the variable measuring the effect of the treatment (ban) on hate speech usage, computed using Difference-in-Differences regression analysis. We ran permutation tests to obtain one-sided p-values.

Usage of manually filtered hate words

Group	**diff_{pre}**	**$\text{diff}_{\text{post}}$**	**Δ**	**DiD coef**	**p-value**
FPH users	0.0013	0.0001	**-90.63%**	-0.0012	0.034
CT users	0.0040	0.0008	**-81.08%**	-0.0043	0.001

Usage of automatically generated hate words

Group	**diff_{pre}**	**$\text{diff}_{\text{post}}$**	**Δ**	**DiD coef**	**p-value**
FPH users	0.0070	0.0013	**-81.99%**	-0.0057	0.038
CT users	0.0117	0.0048	**-59.06%**	-0.0090	0.001

$$y_{ts} \sim x_s + d_{ts} + a_t, \tag{1}$$

where y_{ts} is the amount of hate speech in subreddit s at time t, x_s indicates whether s is one of the treatment subreddits, a_t is an indicator for time t, and $d_{ts} = \delta(t > \tau) \cdot x_s$ where $\delta(t > \tau)$ indicates whether time t is after the ban (at time τ). The coefficient on d_{ts} represents the effect of the ban in this DiD model.

In our work, the DiD analysis calculates the effect of the treatment (i.e., $d_{ts} = 1$ or independent variable) on the outcome (i.e., hate speech usage or dependent variable) by comparing the average change over time in the outcome variable for the treatment group, compared to the average change over time for the control group. The results of the DiD analysis are shown in Table 4, and the effect is measured by the coefficient of d_{ts} (DiD coef).[11] The results in Table 4 demonstrate a dramatic decrease in hate speech usage by the treatment users post-ban. The pre-ban use of manually filtered hate words by the r/CoonTown users ranged between 0.3% and 0.6%, therefore a coefficient of -0.4% represents a large drop in the overall hate speech usage. The pre-ban use of manually filtered hate words by the r/fatpeoplehate users ranged between 0.05% and 0.2%, therefore a coefficient of -0.1% also represents a meaningful drop in the overall hate speech usage.

In order to establish that the apparent effects of the ban were not due to chance, we again performed permutation tests [38]. A detailed description of the procedure and interpretation of the corresponding p-values can be found in *Appendix A*. For r/CoonTown, a one-sided significance test gives $p \approx 0.01$, indicating strong evidence that the drop in hate speech (measured by the usage of both automatically generated and manually filtered hate words) was not due to chance. For r/fatpeoplehate, the same test gives $p \approx 0.03$, also indicating strong evidence that the drop in hate speech was not due to chance.

5 RQ2: COMMUNITY-LEVEL EFFECTS OF THE BAN

Next, we explore the community-level effects of the ban, through the following research questions:

[11] This analysis implicitly assumes that the ban did not affect hate speech usage by the control users. However, if the ban increased hate speech usage by the control users, then this analysis will be conservative.

Table 5. Examples of subreddits that were invaded by banned community users (migrants), ordered by the increase in the migrant's posting activity within these subreddits. The posting activity of migrants nearly doubled within these subreddits post-ban, as they reallocated their activity away from r/fatpeoplehate and r/CoonTown.

Invasion by FPH	**Invasion by CT**
RoastMe	hittableFaces
fo4	The_Donald
JustCause	homeland
MrRobot	thelongdark
FieldOfKarmicGlory	BlackCrimeMatters
prowrestling_ja	RoastMe
bladeandsoul	anime_irl
Voat	OpenandHonest
Vermintide	ModelNASCAR
nakedandafraid	FargoTV

RQ2: What effect did the ban have on subreddits that saw an influx of banned subreddit users?

RQ2a: To which subreddits did the contributors to banned subreddits migrate after the ban?

RQ2b: How did the hate speech usage by migrants change in these subreddits, if at all?

RQ2c: How did the hate speech usage by preexisting users change in these subreddits, if at all?

5.1 Overarching Interrupted Time Series Strategy

The causal inference question is whether the banning of a subreddit causes users from the subreddit to post the same hate speech content elsewhere. While we established effects on users participating heavily in r/fatpeoplehate and r/CoonTown in RQ1, the ban also has possible second-order effects: with those spaces removed, r/fatpeoplehate and r/CoonTown users might reallocate their deviant behavior to other subreddits. In other words, other subreddits might inherit the problem. Unlike RQ1 where we performed user-level matching, the unit of analysis in RQ2 is the subreddit. We do not employ a subreddit-matching strategy to identify subreddits similar to those invaded by banned community migrants because there are not enough subreddits to obtain valid matches. Therefore, we use another causal inference strategy—Interrupted Time Series [2]—to measure the causal effects of the ban. Our techniques include: identifying subreddits that inherited many users formerly active in r/fatpeoplehate and r/CoonTown, and using an interrupted time series procedure to compare the pre- and post-differences in hate speech within these subreddits.

5.2 Subreddits Invaded by Treatment Users: RQ2a

We begin by compiling a list of subreddits where treatment users migrated following the ban. First, we examine all posts present in the timelines of treatment users, and obtain all unique subreddits on which treatment users posted. We tabulate the list of subreddits where treatment users migrated post-ban. We focus on subreddits where treatment users posted pre-ban, and their activity in these subreddits increased post-ban (by 100%). These constitute our *invaded subreddits*. Using this method, we identify 1201 subreddits invaded by r/fatpeoplehate migrants and 275 subreddits invaded by r/CoonTown migrants.

Next, we look at the content analysis of posts in invaded subreddits, where we draw insights from the text contained in posts made on these invaded subreddits.

5.3 Results: RQ2a

Examples of invaded subreddits that received increased post-ban participation from treatment users are shown in Table 5. There were also instances of temporary subreddits that were created immediately after the ban, which served as regrouping places, where treatment FPH and CT users coordinated their next steps. But these subreddits were either banned by Reddit, or died out due to inactivity in the few weeks following the ban (e.g., r/fatpeoplehate1, r/fatpeoplehate2, r/itsacondishun, r/wedislikefatpeople, and so on).

5.4 Hate Speech Analysis: RQ2b & RQ2c

We perform content analysis on the text present in subreddit timelines to answer RQ2b and RQ2c. We examine the variations in hate speech usage in invaded subreddits by two groups of users: *migrants* and *preexisting* users. *Migrants* are users from the banned subreddits, who increased their posting activities by at least 100% in the invaded subreddits following the ban. *Preexisting users* are users who post in the invaded subreddits, but were not a part of the banned subreddits. By computing the variations in hate speech usage of migrants, we examine whether these users bring content from r/fatpeoplehate and r/CoonTown into these invaded subreddits. By computing the variations in hate speech usage of preexisting users, we examine whether these users are influenced by the migration of users from r/fatpeoplehate and r/CoonTown. As migrants and preexisting users comprise the whole user population of the invaded subreddits, we can make claims about entire community effects by combining results from the two.

We compute the frequency of occurrence of words present in the hate lexicons that we generated (both automatically-generated and manually-filtered hate words). Similar to the hate speech analysis in RQ1b, we calculate hate speech usage as the frequencies of hate words normalized by the total number of words in all user timeline posts.

5.5 Results: RQ2b and RQ2c

The temporal variations in hate speech usage within the invaded subreddits are shown in Figure 4. We analyzed over 9 million posts from subreddits invaded by migrants from r/CoonTown, and over 25 million posts from subreddits invaded by migrants from r/fatpeoplehate. The hate speech usage within subreddits invaded by r/fatpeoplehate migrants remained relatively unaffected post-migration. Within subreddits invaded by r/CoonTown migrants, we observed an uptick in hate speech usage post-migration (top left, Figure 4). But there were spikes in hate speech usage that existed even before the ban, and these were not caused by the ban (control lines in Figure 4). There is a possibility that this evidence of an upward trend in hate speech usage was due to a general trend and not specifically due to the ban. In order to control for this, we performed an Interrupted Time Series (ITS) regression analysis. This allows us to make claims about the effects of the ban on change in hate speech usage within these subreddits. A detailed description of the procedure and interpretation of the corresponding p-values can be found in *Appendix B*.

The results from the ITS analysis are shown in Table 6. Through the ITS analysis, we observed that the ban caused no significant changes in hate speech usage by migrants (RQ2b) and preexisting users (RQ2c) within the invaded subreddits. For subreddits invaded by r/CoonTown migrants, a one-sided significance test gives p-value≈ 0.36 for migrants and p-value ≈ 0.36 for preexisting users, indicating a lack of evidence that the increase in hate speech (measured by the usage of manually filtered hate words) was caused by the ban. For subreddits invaded by r/fatpeoplehate migrants, a one-sided significance test gives p-value≈ 0.25 for migrants and p-value ≈ 0.44 for preexisting users, indicating a lack of evidence that the increase in hate speech (measured by the

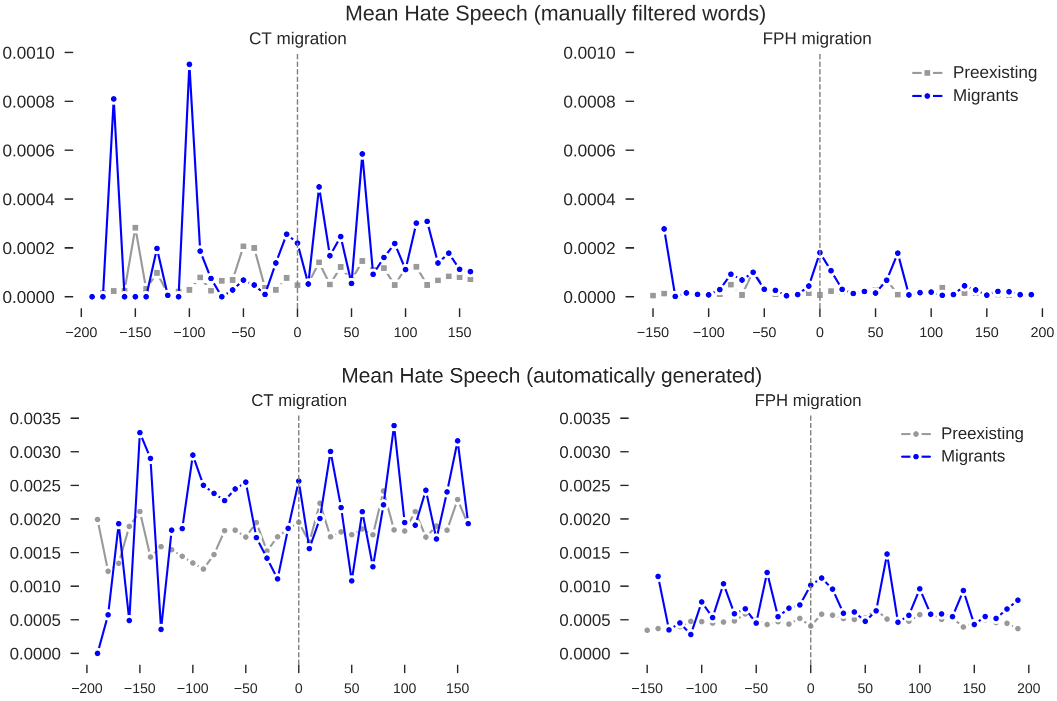

Fig. 4. Variations in hate speech usage on subreddits that received increased activity from treatment users, which we call *invaded subreddits*. We compute mean hate speech usage in all posts made in 2015, obtained from the invaded subreddits. Note that users are only considered if they posted at least five words within invaded subreddits.

usage of manually filtered hate words) was caused by the ban. We see similar results when using automatically generated hate words to measure hate speech usage within the invaded subreddits.

6 DISCUSSION

We presented a novel empirical approach to study the effectiveness of banning deviant hate groups in online communities, investigating the causal effects of Reddit banning r/fatpeoplehate and r/CoonTown in 2015. Next, we reflect on the user-level and community-level effects of Reddit's 2015 ban, discuss the success of the ban in reducing hate speech on the site, and conclude by considering implications for online moderation and online communities more broadly.

6.1 RQ1: User-level Effects of the Ban

Following Reddit's 2015 ban, a large, significant percentage of treatment users from the banned communities left Reddit, as compared to a cohort of control users (RQ1a).[12] Moreover, the ban initiated a chain of events that led to a decrease in overall activity on Reddit (see control in Figure 2). On July 2, 2015, Reddit fired its Director of Communications, Victoria Taylor, without any notice being given to moderators who depended on her for the operation of r/IAmA, one of the most popular subreddits [29]. This resulted in a major blackout on Reddit, where 2,278 subreddits (some

[12] All subsequent effect sizes discussed in this section are significant at the $p - value < 0.05$ level.

Table 6. Interrupted Time Series Regression results for hate speech usage within invaded subreddits. The β coefficient from ITS regression, as well as the one-tail p-values when using both manually filtered and automatically generated hate words are also included. Overall, results show no causal effect of the ban on invaded subreddit hate speech.

Usage of manually filtered hate words

Group	Invasion by	β	p-value
Migrants	CT	6.875e-05	0.362
Preexisting	CT	-1.923e-05	0.360
Migrants	FPH	2.396e-05	0.247
Preexisting	FPH	2.193e-06	0.440

Usage of automatically generated hate words

Group	Invasion by	β	p-value
Migrants	CT	-7.512e-04	0.122
Preexisting	CT	1.186e-04	0.251
Migrants	FPH	1.288e-04	0.258
Preexisting	FPH	2.708e-04	0.136

of which had millions of members) shut down in protest [26]. Overall, these events resulted in decreased user activity levels on Reddit. A notable advantage of the causal inference approach taken in this paper is that it accounts for site-wide variation at and around the time of the ban.

For the banned community users that remained active, the ban drastically reduced the amount of hate speech they used across Reddit by a large and significant amount (RQ1b). Following the ban, Reddit saw a 90.63% decrease in the usage of manually filtered hate words by r/fatpeoplehate users, and a 81.08% decrease in the usage of manually filtered hate words by r/CoonTown users (relative to their respective control groups). The observed changes in hate speech usage were verified to be caused by the ban and not random chance, via permutation tests (see *Appendix A*).

Though we have evidence that the user accounts became inactive due to the ban, we cannot guarantee that the users of these accounts went away. Our findings indicate that the hate speech usage by the remaining user accounts, previously known to engage in the banned subreddits, dropped drastically due to the ban. This demonstrates the effectiveness of Reddit's banning of r/fatpeoplehate and r/CoonTown in reducing hate speech usage by members of these subreddits. In other words, even if every one of these users, who previously engaged in hate speech usage, stop doing so but have separate "non-hate" accounts that they keep open after the ban, the overall amount of hate speech usage on Reddit has still dropped significantly.

6.2 RQ2: Community-level Effects of the Ban

Following the banning of r/fatpeoplehate and r/CoonTown, the affected users migrated to other parts of Reddit. The majority of r/CoonTown users migrated to other subreddits (like r/The_Donald, r/homeland, r/BlackCrimeMatters) where racist behavior has either been noted or is prevalent. On the other hand, most of the r/fatpeoplehate users migrated to qualitatively different subreddits dedicated to roasting users who voluntarily post pictures of themselves or others (r/RoastMe), gaming (r/fo4) or TV shows (r/MrRobot).

We observed no change in the hate speech usage of migrants in the invaded subreddits post-ban (p-value$\geq$ 0.122; the lower-bound in Table 6), nor did we see any significant change in the hate speech usage of preexisting users in these subreddits (p-value$\geq$ 0.136). In simpler terms, the migrants did not bring hate speech with them to their new communities, nor did the longtime residents pick it up from them. Reddit did not "spread the infection."

6.3 Banning Subreddits Worked for Reddit

For the definition of “work” framed by our research questions, the ban *worked for Reddit.* It succeeded at both a user level and a community level. Through the banning of subreddits which engaged in racism and fat-shaming, Reddit was able to reduce the prevalence of such behavior on the site. The amount of hate speech generated across Reddit by treatment users went down drastically following the ban. By shutting down these echo chambers of hate, Reddit caused the people participating to either leave the site or dramatically change their linguistic behavior (as measured via our hate lexicons).

At a community-level, the ban also worked. The subreddits that inherited the activity of former r/fatpeoplehate and r/CoonTown users did not inherit their previous behavior. We examined users’ hate speech usage post-ban in new subreddits—hate speech previously largely confined to these two banned subreddits. When controlling for general trends in hate speech usage across Reddit, we found that the ban had no effect on the hate speech usage in subreddits invaded by migrants from r/fatpeoplehate.[13]

6.4 Possible Reasons Behind the Ban’s Effectiveness

While our approach implies a causal link between the ban and subsequent reductions in hate speech, it does not lay bare *why the ban worked.* Next, we share some of our early thoughts on why the ban worked to control hate speech on Reddit.

We know that the banning of the subreddits r/fatpeoplehate and r/CoonTown led to a dispersal migration to other parts of the site. Yet, the former r/fatpeoplehate and r/CoonTown users could not find other outlets to engage in similar behavior following the ban. As a result, the preexisting subreddits that inherited the activity of former r/fatpeoplehate and r/CoonTown users saw no significant changes in hate speech usage following the ban. This is the finding from RQ2. Perhaps existing community norms and moderation policies within these other, well-established subreddits prevented the migrating users from repeating the same hateful behavior. We have heard anecdotal accounts of this from some Redditors—notably from some members of r/KotakuInAction.

We also know that just after the ban, many temporary r/fatpeoplehate and r/CoonTown variants came into existence (e.g., r/fatpeoplehate1, r/fatpeoplehate2, r/itsacondishun, r/wedislikefatpeople, etc.); those were also banned by Reddit, before they could attain critical mass. Reddit’s strategy of banning copycat subreddits could have also encouraged other moderators to stamp out this type of behavior for fear of running afoul of site administrators. Anecdotally, we have heard that subreddits and their members consciously made efforts to not attract the attention of Reddit site administrators around the time of the ban, fearing their subreddits might be next.

Furthermore, given that former r/fatpeoplehate and r/CoonTown users were unable to find suitable alternatives on Reddit, the ban may have led to the migration of power users from these subreddits to other parts of the Internet. One possible explanation for this is that it is simply easier: instead of constantly hiding from Reddit’s admins, find a new host site. Prior work has found that many Reddit users migrated to other sites following the 2015 bans, where they regrouped (e.g., v/fatpeoplehate and v/[n-word] on Voat) [29]. As a result, the hate speech generated by former r/fatpeoplehate and r/CoonTown users, who were previously active on Reddit, dropped when they abandoned their accounts and left the site. Therefore the migration of users to other sites could have also played a role in reducing the amount of hate speech generated within Reddit, following the subreddit bans.

[13] The code used for our analysis can be found at: https://bitbucket.org/ceshwar/reddit_2015_bans

6.5 Implications for Online Moderation

Reddit's decision to ban r/fatpeoplehate and r/CoonTown—and thereby disperse participants to other parts of the site—reduced overall hate speech usage on the site. An implication for sites is that banning the spaces where deviant groups congregate is likely to work. However, *whether* to ban groups for engaging in a behavior the site considers deviant is a difficult and open question. To start with, who gets to define "deviant?" We have focused here on the pragmatic effects of Reddit's decision to ban r/fatpeoplehate and r/CoonTown, rather than if Reddit should have done it in the first place.

Ideas around freedom of speech online—and conversely, a platform's responsibility to protect its users, community and brand from harm—are undergoing rapid negotiation (e.g., [44]). Some argue for nearly unrestricted freedom of speech on the internet, even surpassing what the most permissive liberal democracies allow. And yet, the platforms are usually owned by companies that have a financial stake in the ongoing success of the platform, as well as no obligation to uphold freedom of speech guarantees. The argument is complex and multi-faceted, with many social, legal and technical layers. For the foreseeable future, however, moderation and banning seem likely to remain in the toolbox for social platforms. The empirical work in this paper suggests that when narrowly applied to small, specific groups, banning deviant hate groups can work to reduce and contain the behavior. We would argue that the efficacy of these strategies should inform conversations around their possible future use.

6.6 Implications for Other Online Communities

Recent work has shown that some banned subreddit users migrated to other social media sites like Voat, Snapzu, and Empeopled [29]. The banning of r/fatpeoplehate and r/CoonTown led to the rise of alternatives on *Voat.co*, for example, where the core group of users from Reddit reorganized. For instance, in another ongoing study, we observed that 1,536 r/fatpeoplehate users have exact match usernames on *Voat.co*. The users of the Voat equivalents of the two banned subreddits continue to engage in racism and fat-shaming [22, 45].

In a sense, Reddit has made these users (from banned subreddits) *someone else's problem*. To be clear, from a macro persepctive, Reddit's actions likely did not make the internet safer or less hateful. One possible interpretation, given the evidence at hand, is that the ban drove the users from these banned subreddits to darker corners of the internet.

7 LIMITATIONS & FUTURE WORK

While we find these results encouraging, they raise a number of questions, challenges and issues. Here, we reflect on some of the limitations present in our work, with an eye toward how we and others might build upon it.

Two communities. Our current work is limited to users in two of Reddit's most prominent hate communities: r/fatpeoplehate and r/CoonTown. Though important, there are still many hate communities on Reddit that we have not explored. There is an opportunity to extend this work to other hate and deviant communities.

Generated hate lexicon. The empirical part of our work examines the usage of hate words generated from posts made in two communities, namely r/fatpeoplehate and r/CoonTown. We have not considered pieces of data including other subreddits that engaged in hate speech, existing blacklists, and so on. Future work exploring and leveraging other pieces of data could paint a richer picture of Reddit's hate communities, and possibly aid in wider coverage of hate speech.

Reasons for account termination. Note that we do not know the exact date at which a Reddit user account was abandoned, nor the exact reason behind the termination of an account. For instance, it could have been the case that a particular account was a “throwaway" used temporarily by a user [25]. We do not account for such things in our current work, and future work may consider the reasons behind account terminations.

8 CONCLUSION

In this paper, we studied the 2015 ban of two hate communities on Reddit, r/fatpeoplehate and r/CoonTown. Looking at the causal effects of the ban on both participating users and affected communities, we found that the ban served a number of useful purposes for Reddit. Users participating in the banned subreddits either left the site or (for those who remained) dramatically reduced their hate speech usage. Communities that inherited the displaced activity of these users did not suffer from an increase in hate speech. While the philosophical issues surrounding moderation (and banning specifically) are complex, the present work seeks to inform the discussion with results on the efficacy of banning deviant hate groups from internet platforms.

9 APPENDIX A: PERMUTATION TESTS (RQ1)

In order to determine whether the pre-ban/post-ban changes in behavior could be explained by chance, we calculated p-values for many of the key estimates in this paper. For many of these estimates (RQ1a, RQ2b, RQ2c) we used classical p-values based on asymptotic results.

9.1 Change in Hate Speech Usage: RQ1b

Because of potential clustering at the subreddit level, for RQ1b we developed p-value using a permutation test [38]. To conduct this test, we permuted the treatment and control labels between the actual treated subreddits (r/fatpeoplehate and r/CoonTown) and the control subreddits. We ran multiple simulations where randomly selected pairs of “highly likely to be banned” subreddits (shown in Table 2) were hypothetically considered to be banned in June 2015. Depending on the subreddits considered to be banned in a simulation, all users who posted in these subreddits (pre-ban) were assigned to the treatment group. All other users were assigned to the control group. For each simulation, we repeated the DiD analysis and obtained a new value of the coefficient for the $\mathrm{d_{ts}}$ variable, DiD coef. After running this simulation many times, we obtained a distribution of coefficients for the d_{ts} variable, DiD coef. This distribution is then compared to the actual DiD coef from our analysis. If the actual DiD coef is “extreme” in relation to the distribution, then this is evidence that the effect estimate we observe was unlikely to have been produced by chance.

To formalize the notion of “extreme” we use p-value, which denotes the proportion of the distribution that exceeds the actual value. For our analysis, because we are interested in drops in hate speech, the p is the proportion of the distribution that is smaller than the actual DiD coef. Because the distribution is constructed by simulation, we set up the simulations to terminate when the p-value of the distribution of DiD coef's obtained reach saturation (i.e., no change in p-value following consecutive simulations).

9.2 Change in Account Activity and Posting Volume: RQ1a

We conducted similar permutation tests to verify the causal effects of the ban on changes in posting volumes and account inactivity/deletions (RQ1a). Change in account activity was measured using the ratios of inactive and deleted accounts among treatment and control users. Change in posting volume was measured through DiD regression analysis, similar to the DiD analysis for change in hate speech usage (described in RQ1b).

10 APPENDIX B: INTERRUPTED TIME SERIES (RQ2)

In RQ2 we are studying the effects of the ban on both the migrants and the preexisting users of the invaded subreddits, therefore, we employ an interrupted time series (ITS) analysis instead of a DiD strategy. In an ITS analysis, hate speech behavior for a group (either migrants or preexisting) is tracked over time, and an regression is used to determine if a treatment at a particular point in time (the ban in this case) caused a change in the behavior (i.e., interrupted the time series). The basic idea of an ITS regression is to model the behavior of the time series before the treatment is applied in order to predict how the series would have looked had the treatment not been applied. This is most often done by regressing the outcome data on time while also including in the regression an indicator variable for the post-treatment time periods. If the analyst were to simply compare the averages of the outcome pre- and post-treatment without regressing the outcome also on time, then one could easily misinterpret a steadily increasing (or decreasing) time series as a treatment effect. In this sense an ITS regression shows that the behavior actually changed at the treatment time, instead of simply a general trend, which would could appear to be attributed to any randomly chosen time.

10.1 Results from ITS Analysis

The results from the ITS study can be seen in Table 6. We observed that there was no significant evidence of change in hate speech among the migrants (RQ2b) and preexisting users (RQ2c) within the invaded subreddits. Through the ITS regression analysis on the usage of manually filtered hate words, we obtained a one-sided p-value = 0.36 for migrants from r/CoonTown and one-sided p-value= 0.25 for migrants from r/fatpeoplehate, indicating a lack of evidence that their increase in hate speech usage within subreddits they migrated to post-ban, was significant and caused by the ban. These p-values were obtained using the results from standard regression software.[14] Informally, this lack of evidence can be seen in Figure 4, where there does not appear to be an obvious change in the trends at the time of the ban.

Using another ITS analysis, we also see no significant changes in hate speech usage for preexisting users within the subreddits invaded by migrants from r/CoonTown or r/fatpeoplehate. Similar results are obtained when using automatically generated hate words to measure hate speech usage within the invaded subreddits.

11 ACKNOWLEDGMENTS

We thank the social computing group at Georgia Tech for their valuable inputs that improved this work. Chandrasekharan and Gilbert were supported by the National Science Foundation under grant IIS-1553376. Eisenstein and Pavalanathan were supported by the National Science Foundation under award RI-1452443, by the National Institutes of Health under award number R01GM112697-01, and by the Air Force Office of Scientific Research.

REFERENCES

[1] Alberto Abadie. 2005. Semiparametric difference-in-differences estimators. *The Review of Economic Studies* 72, 1 (2005), 1–19.

[2] James Lopez Bernal, Steven Cummins, and Antonio Gasparrini. 2017. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology* 46, 1 (2017), 348–355.

[3] Michael S Bernstein, Andrés Monroy-hernández, Drew Harry, Paul André Katrina Panovich, and Greg Vargas. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *In Proc. Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11).*

[14] Because this analysis assumes independence across time, these p-values are likely too small and hence there may even be less evidence for an effect of the ban than is implied by the p-values.

[4] Sam Biddle. 2015. Reddit (Finally) Bans CoonTown. http://gawker.com/reddit-finally-bans-coontown-1722332877. *Gawker* (2015).

[5] Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making. In *Internet, Policy and Politics Conference, Oxford, United Kingdom.*

[6] Stevie Chancellor, Zhiyuan Jerry Lin, and Munmun De Choudhury. 2016. "This Post Will Just Get Taken Down": Characterizing Removed Pro-Eating Disorder Social Media Content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, 1157–1162.

[7] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, 1201–1213.

[8] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The Bag of Communities: Identifying Abusive Behavior Online with Preexisting Internet Data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM.

[9] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. In *Eighth International AAAI Conference on Web and Social Media.*

[10] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Ninth International AAAI Conference on Web and Social Media.*

[11] Tiago Oliveira Cunha, Ingmar Weber, Hamed Haddadi, and Gisele L Pappa. 2016. The Effect of Social Feedback in a Reddit Weight Loss Community. In *Proceedings of the 6th International Conference on Digital Health Conference.* ACM.

[12] Munmun De Choudhury, Sanket Sharma, and Emre Kiciman. 2016. Characterizing dietary choices, nutrition, and language in food deserts via social media. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* ACM, 1157–1170.

[13] Venita DePuy, Vance W Berger, and YanYan Zhou. 2005. Wilcoxon–Mann–Whitney test. *Encyclopedia of statistics in behavioral science* (2005).

[14] Caitlin Dewey. June, 10, 2015. These are the 5 subreddits Reddit banned under its game-changing anti-harassment policy - and why it banned them. https://www.washingtonpost.com/news/the-intersect/wp/2015/06/10/these-are-the-5-subreddits-reddit-banned-under-its-game-changing-anti-harassment-policy-and-why-it-banned-them/. *The Washington Post* (June, 10, 2015).

[15] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying.. In *The Social Mobile Web.* 11–17.

[16] Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.* 182–188.

[17] Michelle Broder Van Dyke. 2015. Reddit Users Revolt After Site Bans "Fat People Hate" And Other Communities. https://www.buzzfeed.com/mbvd/reddit-users-revolt-after-site-bans-fat-people-hate-and-othe. *Buzzfeed News* (2015).

[18] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning.* Omnipress, 1041–1048.

[19] Constantine E Frangakis and Donald B Rubin. 2002. Principal stratification in causal inference. *Biometrics* 58, 1 (2002), 21–29.

[20] Keegan Hankes. 2015. Black Hole. https://www.splcenter.org/fighting-hate/intelligence-report/2015/black-hole. *SPLC Intelligence Report* (2015).

[21] Sara Kiesler, Robert Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* (2012).

[22] Jason Koebler. June, 10, 2015. This Is the Site Redditors Are Migrating to Now That r/fatpeoplehate Is Banned. https://tinyurl.com/voat-fph. *Motherboard* (June, 10, 2015).

[23] Irene Kwok and Yuzhou Wang. 2013. Locate the hate: detecting tweets against blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence.* AAAI Press, 1621–1622.

[24] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 543–550.

[25] Alex Leavitt. 2015. This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing.* ACM, 317–327.

[26] J. N. Matias. 2015. What just happened on reddit? Understanding the moderator blackout. http://socialmediacollective.org/2015/07/09/what-justhappened-on-reddit-understanding-the-moderatorblackout/. *The Washington Post* (2015).

[27] Aiden R McGillicuddy, Jean-Gregoire Bernard, and Jocelyn Ann Cranefield. 2016. Controlling Bad Behavior in Online Communities: An Examination of Moderation Work. (2016).

[28] Justin Wm. Moyer. 2015. 'Coontown': A noxious, racist corner of Reddit survives recent purge. https://www.washingtonpost.com/news/morning-mix/wp/2015/07/17/coontown-a-noxious-racist-corner-of-reddit-survives-recent-purge/. *The Washington Post* (2015).

[29] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Tenth International AAAI Conference on Web and Social Media*.

[30] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 145–153.

[31] Jessica Annette Pater, Yacin Nadji, Elizabeth D Mynatt, and Amy S Bruckman. 2014. Just awful enough: the functional dysfunction of the something awful forums. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2407–2410.

[32] J. Michael Paul and Mark Dredze. 2013. Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 168–178. http://aclweb.org/anthology/N13-1017

[33] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and Consequences in Geotagged Twitter Data. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2138–2148. https://doi.org/10.18653/v1/D15-1256

[34] Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*. Springer, 16–27.

[35] Adi Robertson. 2015. Reddit announces new anti-harassment rules. http://www.theverge.com/2015/5/14/8606923/reddit-anti-harassment-policy. *The Verge* (2015).

[36] Adi Robertson. 2015. Reddit bans 'Fat People Hat' and other subreddits under new harassment rules. http://www.theverge.com/2015/6/10/8761763/reddit-harassment-ban-fat-people-hate-subreddit. *The Verge* (2015).

[37] Daniel M Romero, Brian Uzzi, and Jon Kleinberg. 2016. Social Networks Under Stress. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 9–20.

[38] Paul R Rosenbaum. 2002. Observational studies. In *Observational Studies*. Springer, 1–17.

[39] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[40] Donald B Rubin and Elizabeth A Stuart. 2006. Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics* (2006), 1814–1826.

[41] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 111–125.

[42] Yanchuan Sim, L. Brice D. Acree, H. Justin Gross, and A. Noah Smith. 2013. Measuring Ideological Proportions in Political Speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 91–101. http://aclweb.org/anthology/D13-1010

[43] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology* 63, 2 (2012), 270–285.

[44] Nabiha Syed Syed and Ben Smith. June, 2, 2016. A First Amendment For Social Platforms. https://medium.com/@BuzzFeed/a-first-amendment-for-social-platforms-202c0eab7054. *Medium* (June, 2, 2016).

[45] Voat. 2015. "This is the official r/CoonTown replacement". https://tinyurl.com/voat-ct-replacement. (2015).

[46] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 19–26.

[47] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 88–93.

[48] Anne Weber. 2009. *Manual on hate speech*. Council of Europe.

[49] Jun-Ming Xu, Benjamin Burchfiel, Xiaojin Zhu, and Amy Bellmore. 2013. An Examination of Regret in Bullying Tweets.. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 697–702.

[50] Zhi Xu and Sencun Zhu. 2010. Filtering offensive language in online communities using grammatical relations. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*. 1–10.

[51] George Kingsley Zipf. 1949. Human behavior and the principle of least effort. (1949).

Received June 2017; revised August 2017; accepted November 2017