Polit Behav
DOI 10.1007/s11109-016-9373-5

ORIGINAL PAPER

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

Abstract I conduct an experiment which examines the impact of group norm promotion and social sanctioning on racist online harassment. Racist online harassment de-mobilizes the minorities it targets, and the open, unopposed expression of racism in a public forum can legitimize racist viewpoints and prime ethnocentrism. I employ an intervention designed to reduce the use of anti-black racist slurs by white men on Twitter. I collect a sample of Twitter users who have harassed other users and use accounts I control ("bots") to sanction the harassers. By varying the identity of the bots between in-group (white man) and out-group (black man) and by varying the number of Twitter followers each bot has, I find that subjects who were sanctioned by a high-follower white male significantly reduced their use of a racist slur. This paper extends findings from lab experiments to a naturalistic setting using an objective, behavioral outcome measure and a continuous 2-month data collection period. This represents an advance in the study of prejudiced behavior.

Keywords Online harassment · Social media · Randomized field experiment · Social identity

Electronic supplementary material The online version of this article (doi:10.1007/s11109-016-9373-5) contains supplementary material, which is available to authorized users.

Replication materials are available on the author's website, www.kevinmunger.com.

✉ Kevin Munger
km2713@nyu.edu

[1] Department of Politics, New York University, 19 West 4th Street, 2nd floor, New York, NY, USA

Published online: 11 November 2016

Introduction

The explicit expression of hostile prejudice is no longer acceptable in mainstream US society. This is evidence for changing social norms, though these new norms are not as well-established in some communities, especially on the internet. The rise of online social interaction has brought with it new opportunities for individuals to express their prejudices and engage in verbal harassment.

This behavior has implications for both the perpetrators and their victims. Minorities and other vulnerable populations are frequently the subject of online harassment on social media sites, often in response to expressing views that harassers disagree with (Kennedy and Taylor 2010; Mantilla 2013). They are likely to become more anxious for their safety, more fearful of crime and less likely to express themselves publicly (Henson et al. 2013), systematically de-mobilizing the populations who tend to be victimized (Hinduja and Patchin 2007). Engaging in harassment of non-whites also fuels ethnocentrism among whites, which has been shown to affect how whites feel about political topics like healthcare and immigration (Banks 2014, 2016), and to affect voting outcomes (Kam and Kinder 2012).

Severe online harassment takes the form of explicit threats or the posting of personal information, forcing targets to modify their behavior out of fear for their immediate safety. Although all harassment can contribute to a toxic online community, this paper is specifically about racist harassment of white men against blacks.

There have been many efforts to reduce online harassment on the part of online forums for social interaction, as well as by brick-and-mortar institutions like schools, universities and government agencies. They tend to involve blanket bans on certain behaviors, enforced either through the public promotion of norms or individual sanctions for clear violations enforced by moderators. A comprehensive review of the literature on prejudice reduction and harassment prevention (Paluck and Green 2009) finds that very little of the research in this area is causally well-identified, and calls for more experimental research. I conducted a novel randomized field experiment that is able to measure the causal effect of specific interventions on the real-world harassing behavior of Twitter users, continuously and over time.

I searched for tweets containing a powerful racial slur ("n****r") to identify harassers with public Twitter accounts, and I assigned each subject to the control or to one of four treatment conditions. Using Twitter accounts that I controlled ("bots"), I tweeted at the subjects to tell them that their behavior was unacceptable. I varied two aspects of the bots, resulting in a 2×2 experimental design: the first dimension of variation was the identity of the bot, to test the finding from Social Identity Theory that sanctioning by members of a person's in-group is more effective (Tajfel and Turner 1979). The second variation was in the number of followers the bot had. This tested the "influentials hypothesis", that influential individuals are crucial for driving changes in norms of behavior in society (Aral and Walker 2012).

I also expected to find heterogeneous treatment effects in the degree of anonymity of the subjects. Based on findings from related online contexts (Omernick and Sood 2013), my hypothesis was that more anonymous individuals would be less likely to respond to the treatment. An alternative hypothesis, based on the social identity model of

deindividuation effects (SIDE), would be that more anonymous individuals would actually be more susceptible to this normative pressure (Postmes et al. 2001).

I find support for the hypothesis that the same message had disparate impact based on in-group identity (here, race), with messages sent by white men causing the largest reduction in offensive behavior among a subject pool of white men.[1] However, this effect was only found among messages sent by accounts that had a high number of Twitter followers. This effect persisted for a full month after the application of the treatment. This finding concords with my hypothesis that the largest treatment effect would be that of receiving a message from a high-status white man. However, the effect of the followers treatment and the group identity treatment were multiplicative, rather than additive, as none of the other treatment conditions caused a significant behavioral change.

The results varied by the degree of anonymity of the subjects. The main effect was substantively similar among the anonymous subgroup. Among the subjects who provided some amount of identifying information, though, the reduction disappeared, and there was actually an increase in racist harassment among the subjects who received a message sent by a black bot with few followers. This finding was contrary to my hypothesis, and lends support to the role of anonymity in the SIDE model in this context.

The net effect of all of the treatments in this study was to reduce the rate of racist harassment. Overall, my intervention caused the 50 subjects in the most effective treatment condition to tweet the word "n****r" an estimated 186 fewer times in the month after treatment.

Reducing Manifestations of Prejudice

Racism, which is a necessary component of the racist harassment studied here, is a form of prejudice, which Dovidio and Gaertner (1999) define as an "unfair negative attitude toward a social group or a member of that group", and Crandall et al. (2002) define as "a negative evaluation of a group or of an individual on the basis of group membership." This paper makes the assumption that directing the word "n****r" at another person constitutes racist harassment, regardless of how justified the user believes their prejudice to be.

Beginning with Allport (1954)'s influential work on prejudice, the subject has been well-studied in psychology. Allport's "contact hypothesis"—that mere contact between different groups helps to reduce prejudice that each holds towards the other—has proven difficult to verify causally. A comprehensive review finds only mild support for the contact hypothesis (Pettigrew and Tropp 2006), and others note that the subject makes isolating causation difficult (Binder et al. 2009).

A more promising approach for analyzing the formation and reduction of prejudices has to do with social norms. Group norm theory holds that "social norms [including prejudices] are formed in group situations and subsequently serve as standards for the individual's perception and judgment when he is not in the group

[1] All hypotheses were pre-registered at EGAP.org (ID number 20150520AA) prior to any data collection.

situation" (Sherif and Sherif 1953). Attitudes towards out-groups are a particularly important set of group norms, and prejudice towards out-groups can be a strong signal of in-group membership (Brewer 1999).

Recent experiments have aimed to test the role of group norms in prejudice formation. Prejudiced attitudes can be reduced (in the short term) by priming less prejudiced social identities; by increasing individual salience vis-a-vis group membership; and by using a confederate to challenge people's understanding of group norms (Plant and Devine 1998; Dovidio and Gaertner 1999; Blanchard et al. 1994). These papers, and others in the literature, suffer from a limitation common to experiments run with convenience samples: they cannot track either long-term or non-lab manifestations of prejudice. Two exceptions to the former problem are Stangor et al. (2001), who show that providing consensus information about in-group norms of prejudiced attitudes can affect survey responses a week later; and Zitek and Hebl (2007), who find that social pressure is more effective at changing prejudiced attitudes if the norms are less clear (eg prejudice against obese people) up to a month after the experiment. By studying the behavior of people on Twitter, my approach is able to capture a continuous measure of prejudice reduction over time and in a naturalistic setting.

Although openly harassing people based on their race is not as common now as it once was, online racist harassment is an increasingly large problem. Studies of computer mediated communication (CMC) have some insight as to why: CMC tends to result in less success in applying normative pressure (Kiesler et al. 1984; Walther 1996; Bordia 1997).

The primary mechanism used to explain the differences in CMC over the internet has been postulated to be deindividuation: people become immersed in the medium of discussion and lose a sense of self-awareness. This mechanism is best explained by the SIDE, in which the depressed sense of one's personal identity is supplanted by an increased sense of one's social identity (Reicher et al. 1995; Lea and Spears 1991).

The anonymity enabled by CMC also leads to more racist harassment online. As Moor (2007) describes anonymous online communities, "people are relatively indistinguishable and their memberships of online discussion groups are far more salient than their personal identities." In communicating online, there are fewer dimensions on which people can identify with a group; speech norms are central.

Prejudiced harassment against out-groups has been used to signal in-group loyalty in the physical world, and it serves the same purpose in online communities. Engaging in prejudiced harassment against out-groups—in this case, blacks—primes ethnocentrism and changes the salience of particular political issues like healthcare (Banks 2014) and immigration (Banks 2016). There is also evidence that the expression of prejudiced views online has implications for vote choice, with the most prominent example being the 2008 presidential election. Increased belief in racial stereotypes decreased Barack Obama's vote total (Piston 2010; Kam and Kinder 2012).

But SIDE also suggests an avenue for reducing online racist harassment: individuals' social identities are actually composed of several overlapping identities. It follows that the influence of specific online communities with norms

of online harassment can be diminished by appealing to their other, offline identities. Rather than leading to increased self-regulation and decreased responsiveness to normative pressure, as in classical models of deindividuation, SIDE posits that deindividuation—when enabled by anonymity—should lead to increased response to normative pressure (Postmes et al. 2001).

Still, as Paluck and Green (2009)'s summary of the literature points out, there has been little research done in the field of prejudice reduction using randomized experiments outside of the laboratory. This paper attempts to address this lacuna. It also represents, with Coppock et al. (2015), one of the first randomized control experiments to be conducted entirely on Twitter.

The crucial advantage of this experimental design is that I could measure real behavior continuously for months. In order to quantify this behavior, I operationalized racist online harassment in the form of the use of the word "n****r." This slur is the most substantively important vehicle for racist harassment, and filtering on its use was the fastest way to collect a sample of genuine harassers. I acquired this data by scraping the Twitter history of each subject before and after being treated.

There is a sizable body of research that indicates that attempts to reduce prejudiced behavior are more effective when made by members of the in-group (Rasinski and Czopp 2010; Gulker et al. 2013). There is also evidence that prejudice-reducing efforts made by higher-status individuals are more effective, although the exact definition of "high status" depends on the context. Paluck et al. (2016) find this to be the case when the high status individuals are "social referents" (who other students look to) in a high school, and Shepherd and Paluck (2015) call highly-connected male high schoolers "high status". Aral and Walker (2012) observe differences in peer influence with a large-scale study of Facebook users, finding that influence varies with marital status and gender. In all of these contexts, the theoretical expectation is that "high status" individuals have a greater capacity to define group norms, and that observers are more likely to mimic their behavior to try and fit in with their group.

Because Twitter is a semi-anonymous environment, I drew from both the SIDE literature about group norm promotion and the research on highly influential social referents to motivate my hypotheses and related experimental manipulations. Specifically, I varied the identity of the bots applying the treatment. They were either In-group (white men) or Out-group (black men), and either had many followers or few followers. Based on the findings discussed above, my hypothesis was that the largest treatment effect would be from In-group/High Followers bots and that the smallest treatment effect would be from Out-group/Low Followers bots. I hypothesized that the other two treatment conditions would have medium-sized effects:

Hypothesis 1 The ranking of the magnitudes of the decrease in harassment will be:

$$In\,group/High > \begin{matrix} In\,group/Low \\ Out\,group/High \end{matrix} > Out\,group/Low.$$

Previously, the degree of anonymity allowed in an online community has been shown to affect the prevalence of online harassment, with more anonymity being associated with more harassment (Omernick and Sood 2013; Hosseinmardi et al. 2014). Twitter allows users to be anonymous to the extent that their accounts can be entirely divorced from their real-life persona, but many users choose to provide identifying information like that which identifies my bots.

To create an anonymity score, I examined several aspects of each subject's profile: whether they had a profile picture of themselves[2] and whether a given name was present in their username or handle. I used these to create a categorical anonymity score that ranged from 2 (most anonymous) to 0 (least anonymous).

The above findings about online communication suggest that greater anonymity is associated with more harassment and lower-quality communication, but SIDE theory implies that norm promotion should be stronger in anonymous contexts. The idea is that individuals make less of a distinction between themselves and other members of their group, and are thus more likely to follow group norms than their own idiosyncratic preferences (Postmes et al. 2001).

Neither of these strains of research have direct implications for my experimental design. Here, anonymity is a self-selected covariate of each subject, rather than a global characteristic. My expectation was that subjects who elected to share less personal information would be less invested in their online communities, and thus less likely to care about group norms. My findings show that this expectation was mistaken. The opposite turned out to be the case, with the expected treatment effects found only among the anonymous subjects.

Hypothesis 2 The magnitude of the decrease in harassment will negatively covary with the subject's anonymity score.

Experimental Design

Among the most challenging aspects of studying mass behavior on Twitter is the selection of a meaningful sample of Twitter users. In order to ensure that efforts to reduce racist harassment could be measured, it was essential to have a sample of users who engaged in racist harassment in the first place.

There is a large and growing literature on the automatic detection of online harassment (Yin et al. 2009; Chen et al. 2012). The task of discerning genuine harassment from heated argumentation or sarcastic joking is challenging, but the presence of prima facie offensive language makes it far easier. In fact, in corpuses that contain enough strongly offensive language, a simple dictionary of strongly offensive terms outperforms even sophisticated classifiers. The dictionary approach also has the advantage of being rapidly implementable at scale.

The detection of second-person pronouns, to determine at whom the profanity is directed, is a large and easy improvement on naive profanity detection, and the structure of Twitter use makes this kind of analysis straightforward: tweets that

[2] Whether a picture is actually of the subject was impossible to verify perfectly; I included any picture that clearly showed the face of a person who I did not recognize.

begin with an "@[username]" are explicitly targeted at the recipient. To further refine the search for racist online harassment, I created a sample of individuals who tweeted a racial slur ("n****r") at another account.[3] In the racial context of the United States, this term is almost certainly the most intrinsically offensive, and people who use it thus represent a "hard case" for this experimental design—there is no doubt that these people are aware that directly tweeting this term at another person constitutes harassment.

Using the streamR package for R, I scraped the user information (including the most recent 1000 tweets) of anyone who tweeted the word "n****r" at another user. For each of these users, I applied a simple dictionary method to calculate the average number of offensive words per tweet in the text of those tweets to generate an offensiveness score for that user. As Sood et al. (2012) point out, the problem of selecting a list of "offensive" words is challenging, and some previous efforts have used arbitrary external dictionaries.[4] To avoid false positives, I used a much shorter list of swear words and slurs.[5]

I discarded users whose offensiveness score fell below a certain threshold and who were thus not regularly offensive. To determine what this "regularly offensive" threshold should be, I randomly sampled 450 Twitter users whose accounts were at least 6 months old.[6] I calculated the offensiveness score for these users' most recent 400 tweets and set the threshold for inclusion in the experimental sample at the 75th percentile of offensiveness. Substantively, this meant that at least 3% of their tweets had to include an offensive term.

This addressed many problems that could arise from the use of jokes or sarcasm: a dictionary method like searching for ethnic slurs cannot capture any information about the tone of a tweet, but leveraging more data and richer contextual information makes mis-classification less likely.[7]

There were several other restrictions I placed on the sample of users. Because they are the largest and most politically salient demographic engaging in racist

[3] As is recorded in my pre-analyis plan (registered at EGAP, ID number 20150520AA), I had originally intended to perform two similar experiments: one on racist harassment, and the other on misogynist harassment. However, my method was insufficient for generating a large enough sample of misogynist users. For any misogynist slur I tried to use as my search term (bitch, whore, slut), there were far too many people using it as a term of endearment for their friends for me to filter through and find the actual harassment. I plan on figuring out a way to crowdsource this process of manually discerning genuine harassment, but for now, the misogynist harassment experiment is unfeasible. The pre-analysis plan also intended to test two hypotheses about spillover effects on the subjects' networks, but this has thus far proven technically intractable.

[4] Chen et al. (2012), for example, emulates Xu and Zhu (2010) and takes a list of terms from the website www.noswearing.com.

[5] For a full list of terms, see the Online Appendix.

[6] Each Twitter account is assigned a unique numerical user ID based on when they signed up; newer accounts have higher ID's. Not all of the numbers correspond to extant or frequently used accounts, so if I randomly picked one of those numbers, I generated a new random number.

[7] Still, there are many people who believe that they're "joking" when they call a friend a slur. While this is still objectionable behavior, it is different from the kind of targeted prejudiced harassment that is of interest in this paper, so I excluded from the sample any users who appeared to be friends who did not find the slur they were using offensive. This process is inherently subjective, but it usually entailed the users with a long back-and-forth, with slurs interspersed with more obviously friendly terms.

online harassment of blacks, I only included subjects who were white men. This ensured that the in-groups of interest (gender and race) didn't vary among the subjects, and thus that the treatments were the same. This additional control was essential, given the power of the study. I also included anonymous users because there were a large number of such accounts engaging in prejudiced harassment and I had different theoretical expectations about how they would respond to treatment. I recorded the degree of anonymity on a categorical scale from 0 to 2 based on if they included their real name and/or a picture of themselves. To the extent possible, I also excluded minors from the sample. Most users did not provide their exact age, but I removed from the sample any user who gave an indication of being underage or who mentioned high school.

Because the subjects in this experiment were drawn from a specific subsection of the overall population, the criteria for inclusion discussed above are fundamental. Figure 1 provides a visual overview of the sampling procedure.

After I verified that a user met all of the criteria for inclusion, I assigned him to one of the treatment conditions or the control condition, subject to balance constraints.[8] Because this process was time-consuming, and there were a fixed number of potential subjects who met these criteria tweeting at a given time, the subject discovery and vetting took place in several periods. The first wave of subjects was collected from August 5th to August 7th, 2015; the second wave from August 25th to August 26th; the third wave from September 7th to September 11th; and the last wave from September 14th to September 16th. See Fig. 2 for a visual summary.[9] The crucial advantage of this real-time detection was that the time that elapsed between when a user tweeted the slur and when he received the treatment was under 24 hours, adding to the realism of the treatment.

The actual application of the treatment was straightforward. Depending on which condition the subject was assigned to, I rotated through the bots in that condition and tweeted the message:

> @[subject] Hey man, just remember that there are real people who arehurt when you harass them with that kind of language

Because this was an "@"-reply, it was only visible to anyone who clicked on the harassing tweet, and to the subject himself.

The four experimental conditions are summarized in Table 1. I varied the race of the bots in order to test the findings in Rasinski and Czopp (2010) and Gulker et al. (2013) that in-group sanctioning is more effective than out-group sanctioning: in this case, that the effect of a tweet from a white Twitter user would be greater than one from a black Twitter user. The number of followers a Twitter user has is indicative of how influential they are, at least within the context of Twitter, so I

[8] Throughout the assignment process, I matched subjects in each treatment group on their (0–2) anonymity score. They were otherwise randomly assigned.

[9] This process was approved by NYU's Institutional Review Board. These subjects had not given their informed consent to participate in this experiment, but the intervention I applied falls within the "normal expectations" of their user experience on Twitter. The subjects were not debriefed. The benefits to their debriefing would not outweigh the risks to me, the researcher, in providing my personal information to a group of people with a demonstrated propensity for online harassment.

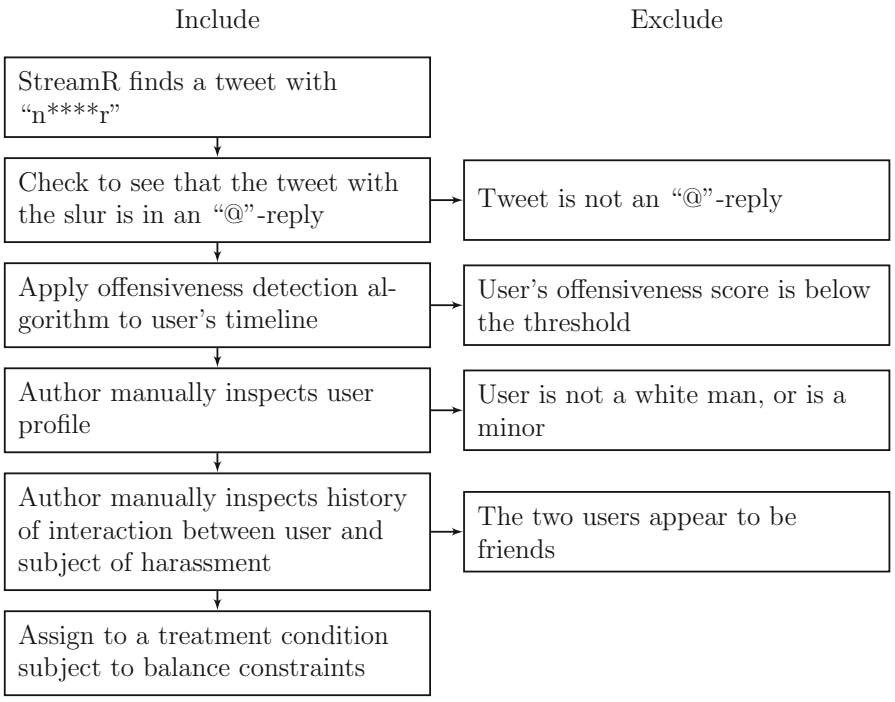

Fig. 1 Sample selection process: This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded

varied that quantity to test the finding in Shepherd and Paluck (2015) and Paluck et al. (2016) that sanctioning by high-status individuals is more effective than that by low-status individuals.

For example, users assigned to the Out-group/Low Followers condition were sent a message like the one seen in Fig. 3a, sent by bot @Rasheed[XXXXXX].[10] After the subject received the treatment, he got a "notification" from Twitter, which caused him to be exposed to the treatment tweet. Because being admonished by a stranger is an uncommon (though far from unknown) experience, the subject was inclined to click on the bots' account; if he did, he saw the bot's profile page, Fig. 3b. @Greg[XXXXXXX] was a bot in the In-group/Low Followers condition. This allowed the subject to clearly determine the race and gender of his admonisher, and to see how many followers the account had (in this case, 2). I could not, however, directly measure this behavior, and it is possible that some subjects did not click on the bot's profile. If that were the case, they would still have noticed the bot's race from the profile picture and username, but they would not have seen the number of followers. This would bias the effect of the followers treatment downward.

As the two bots shown in Fig. 3 illustrate, the variation in the bot identity was accomplished by changing the number of followers, the skin color of the profile picture, username, and full name. To vary the number of followers, I bought

[10] I avoid providing the entire username of the bot to protect my subjects' anonymity.

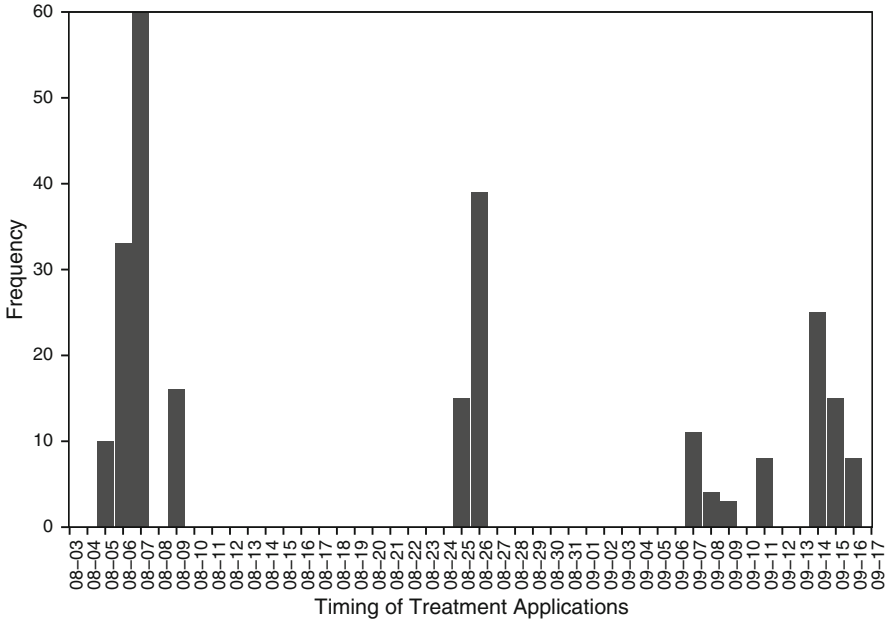

Fig. 2 Timing of the experiment in the field: The number of subjects added to the sample each day is plotted on the *y*-axis. Each treatment was applied within 24 h of the subject tweeting a racial slur. There were potential subjects tweeting every day, but I was only actively searching on the days indicated. All dates 2015

Table 1 Experimental design and hypothesized effect sizes

	In-group	Out-group
Low followers	Medium effect	Small effect
High followers	Large effect	Medium effect

followers for some accounts and not others (Stringhini et al. 2012). In the low-follower condition, the bots had between 0 and 10 followers (some of the bots were followed by other Twitter users, most of them spam accounts). In the high-follower condition, they had between 500 and 550 followers.

When generating the bots, I chose handles that consisted of first and last names that were identifiably male and white or black, following Bertrand and Mullainathan (2003). Because all of these handles were already taken (and Twitter requires that each account have a unique handles), I added random numbers to generate unique handles. The usernames were the first and last name used in the handle without the numbers; usernames do not need to be unique.

The most important aspect of the bots' profile was their profile picture. It was the first thing the subject saw, and was also the largest potential source of bias. In order to maximize the amount of control I had over the treatment, I used cartoon avatars

Fig. 3 Treatments. **a** The treatment—black bot. **b** The bot applying the treatment—white bot

for the profile pictures. If I had used real photos, there would exist the possibility that the particular people pictured varied on some important dimension other than race. This practice does not detract from the verisimilitude of the bot—using cartoon avatars on Twitter is not uncommon. I gave each bot the same facial features and the same professional-looking attire; the only thing I varied was the skin color, using a similar technique to Chhibber and Sekhon (2014).[11]

In order to ensure that the actual treatment experienced by the subject was maximally similar to the "real life" experience of being sanctioned by a stranger on Twitter, it was essential that the subject be unaware that my bot was in fact a bot. If the subject suspected that the bot was not the authentic online manifestation of a concerned citizen, the effects of norm promotion would be attenuated and the measured treatment effect would be a conservative estimate of the true treatment effect. One possible source of skepticism was that the followers I bought were not high-quality followers, in that they were obviously not real accounts; however, having fake or "spam" Twitter followers is not uncommon.

The history of tweets by the bot represented the most serious problem for verisimilitude. Under the "Tweets" tab displayed in Fig. 3b, there needed to be a

[11] It is possible that a stronger racial treatment effect might have obtained if I also changed the facial features of the black bots to be more afrocentric, the effect of which Weaver (2012) finds to be approximately as large as changing skin color on voting outcomes.

plausible history of tweets to convey that this was a real, active user. To that end, I had the bot tweet from a list of personal but innocuous statements ("Strawberry season is in full swing, and I'm loving it") and retweeted a number of generic news articles. However, in the default profile display, tweets that are directed "@" another user are not visible. If the subject clicked on the "Tweets & replies" tab, they became visible, but my innocuous tweets were interspersed so that the treatment tweets represent less than half of the bot's overall tweets. As a result, only three of the 242 subjects responded to accuse my bots of being bots.

Results

The primary outcome of interest was the change in the subjects' levels of offensiveness in the four different treatment arms, relative to the control group. However, I could not collect a full 2 month's worth of tweets for some of the subjects, for one of three reasons: at some point after the treatment, the subject could have made his account private, or he could have deleted his account, or the account could have been banned by Twitter. The first only happened to three accounts out of the 242 in the sample,[12] but I could not distinguish between the last two.[13] Table 2 presents the attrition rates among the different treatment arms in the sample. The average attrition (defined as subjects who dropped out of the sample before tweeting at least 25 times after the treatment) among the four treatment conditions was 16%, compared to 13% among the control subjects, an insignificant difference ($p = 0.58$).[14]

Despite this insignificance, performing the analysis only on the subjects who remained in the sample could introduce post-treatment bias. It is preferable to include all of the subjects, but this requires an assumption about the behavior of the subjects for whom I had missing data. I made the following assumption: for each of these observations with missing post-treatment data, I treated their post-treatment rate of racist language as zero. The subjects who were no longer tweeting publicly had ceased to engage in online harassment.[15]

The results support H_1. In Fig. 4, Panel A shows the effect of the different treatment arms on the absolute daily use of the word "n****r" over the week after

[12] Initially, I assigned 243 subjects to one of the four treatment arms or to the control group. However, the rate of tweeting of one of these subjects was too infrequent for me to be able to calculate a meaningful pre-treatment rate of offensive language use, and I excluded him.

[13] I contacted Twitter to see if they could provide me with this information, but they were not forthcoming.

[14] Note, though, that the Out-group/High Followers condition saw much lower attrition than the other treatment conditions. I have no explanation for why this is the case, and in fact my ex ante expectation was that, to the extent that attrition was positively correlated with any treatment condition, it would have been higher among the High Followers conditions.

[15] A more conservative and less substantively accurate assumption is to treat these observations as having a post-treatment rate of racist language equal to their pre-treatment rate of racist language use. Figure 7 in the Appendix presents the results with this alternate assumption. The results are substantively similar, although the point estimates are slightly smaller.

Table 2 Attrition rates

	Control	In-group Low	Out-group Low	In-group High	Out-group High
Baseline # of subjects	51	49	44	50	48
# with more than 1 post-treatment tweets	49	47	42	47	47
# with more than 25 post-treatment tweets	43	42	38	41	46
Attrition %, fewer than 25 post-treatment tweets	16%	14%	14%	18%	4%

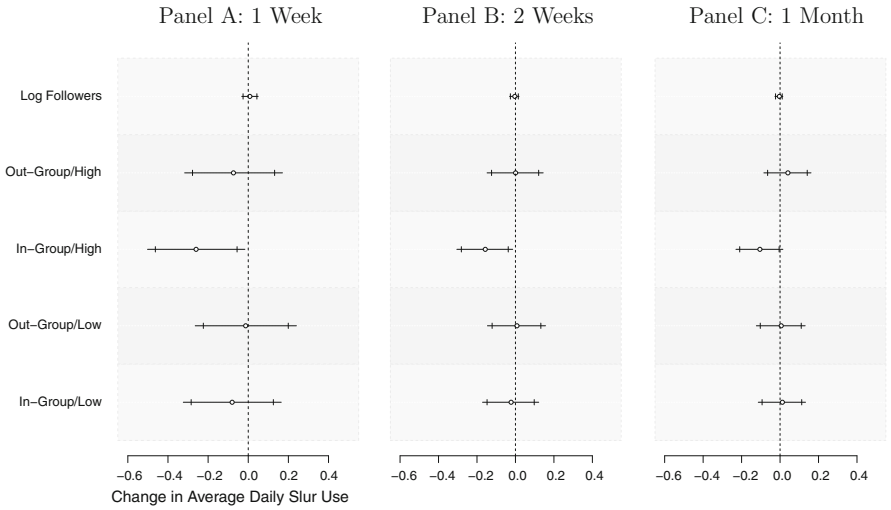

Fig. 4 Full sample ($N = 242$). *Each panel* represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the In-group/High Followers treatment in *Panel A* shows these subjects reduced their average daily usage of this slur by 0.26 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the 2 months prior to the treatment. The *vertical tick marks* represent 90% confidence intervals and the *full lines* represent 95% confidence intervals

the treatment.[16] Panel B expands the time period to 2 weeks, and Panel C expands it to 1 month. Each panel shows the result of an OLS regression in which the dependent variable is the absolute number of instances of racist language during that time period divided by the number of days in that time period. Each regression controls for the subjects' log number of followers, displayed in the first row. Each regression also controls for the average rate of the subjects' use of that offensive term in the 2 months prior to the treatment. The four treatment arms each represent

[16] I have selected my sample based on their use of this slur. Expanding the dependent variable to include other anti-black language does not substantively change the results, primarily because the use of other anti-black slurs is uncommon among this subject pool.

the comparison between that arm and the control group, and each treatment effect is displayed in one of the bottom four rows.

The only treatment that significantly decreased the rate of racist language use was the In-group/High Follower treatment. This is precisely what H_1 predicted to have the largest effect. There is a reduction in racist language use among the other three treatment conditions, but it is not significant at $p < 0.10$, and it is of smaller magnitude than the reduction in the In-group/High Followers condition. This was contrary to my expectations in H_1: I predicted that both the Out-group/High Followers and In-group/Low Followers conditions would have a larger effect than the Out-group/Low Followers condition.

Comparing across the panels of Fig. 4 shows the decay in the effect of the In-group/High Follower over time. Although the effect remains statistically significant, the coefficient decreases steadily. In Panel A, the point estimate of -0.26 indicates that the daily rate of the use of the word "n****r" decreased by 0.26 more among subjects in the In-group/High Follower treatment condition than among subjects in the control condition. This average treatment effect decreased in magnitude to -0.16 in Panel B and -0.11 in Panel C. Treatment effects were not significantly different from zero after 2 months, so these results are not shown.

In order to test H_2, I divide the sample into two subgroups: those with anonymity scores equal to two, indicating that they shared no identifying information, and those with anonymity scores of either zero or one, indicating that they shared their real name, a real picture of themselves, or both. The anonymous sub-group had 158 subjects, and the non-anonymous subgroup had 84. Only 26 subjects had an anonymity score of zero, so I cannot divide this group further. My prediction in H_2 was that the reduction in harassment would be greater among the non-anonymous subgroup.

Figures 5 (anonymous) and 6 (non-anonymous) display the results. Figure 5 roughly mirrors the findings on the entire sample from Fig. 4: there was a significant reduction among the subjects only in the In-group/High Followers condition, although in this case the effect was no longer significant after 1 month. However, the results from Fig. 6 are starkly different. Not only was there no reduction in any treatment condition, there was actually a significant increase in racist language use among subjects in the Out-group/Low Followers condition. This was the condition that H_1 predicted would experience the smallest reduction in racist language use, but the fact that this treatment caused an increase was surprising.

These results not only fail to support H_2, they provide evidence for the opposite conclusion: there was only a decrease in harassment among subjects with the highest anonymity score, and the direction of the effect changed (at least for one treatment arm) among subjects with non-maximal anonymity scores.

Discussion

The primary prediction expressed in H_1, that the In-group/High Follower treatment would cause the largest reduction in racist language use, was borne out. This effect was larger than either the In-group/Low Follower or Out-group/High Follower

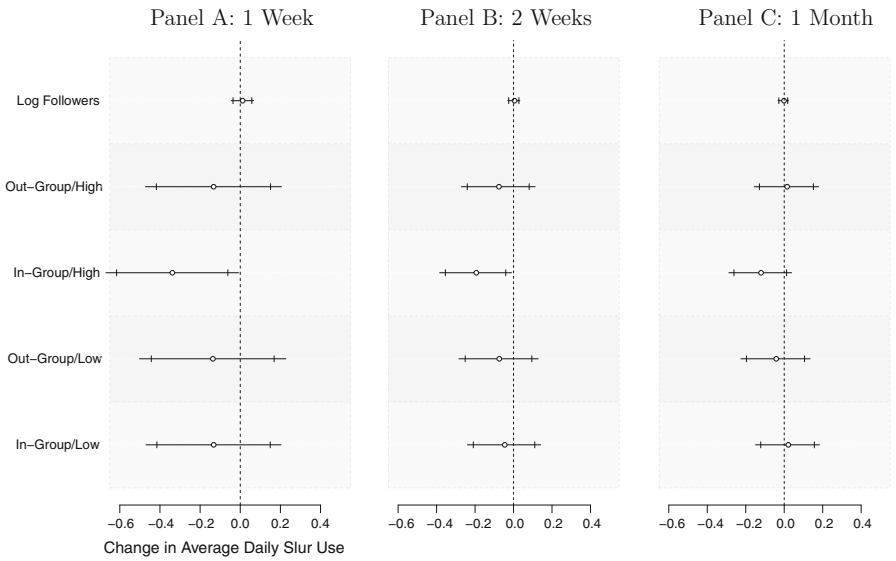

Fig. 5 Anonymous subjects ($N = 158$). *Each panel* represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the In-group/High Followers treatment in *Panel A* shows these subjects reduced their average daily usage of this slur by 0.34 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the 2 months prior to the treatment. The *vertical tick marks* represent 90% confidence intervals and the *full lines* represent 95% confidence intervals

treatments, although these latter two reductions were not significant as expected. Overall, this is evidence of a multiplicative effect of the two treatments, as neither had an effect in isolation.

I found evidence for both social identity theory in terms of in-group norm promotion and the theory that influential community members drive changes in normative group behavior. The sanctioning treatment caused subjects to update their beliefs about norms of online behavior, but only when the sanctioner was both a member of the in-group and perceived to be influential.

Encouragingly, these effects persisted for the first month under study, although not for 2 months. Also, the p value of the effect in the 2 week time period was actually smaller than for the 1 week and 1 month time periods. This non-monotonicity was surprising, relative to my expectation of a steady decay. My post-hoc explanation is that the smaller-than-expected effect sizes in the 1 week time period were caused by some subjects responding directly to the treatment by harassing the bot that tweeted at them and actively rebelling against the attempt to persuade them to change their behavior.

This phenomenon is called "reactance", and it has been shown to occur in a variety of political contexts. In a study of efforts to correct misperceptions, for example, Nyhan and Reifler (2010) find that, when confronted with evidence that a view they hold is false, some people actually become firmer in their false belief. More closely related to the current context, a study by Harrison and Michelson

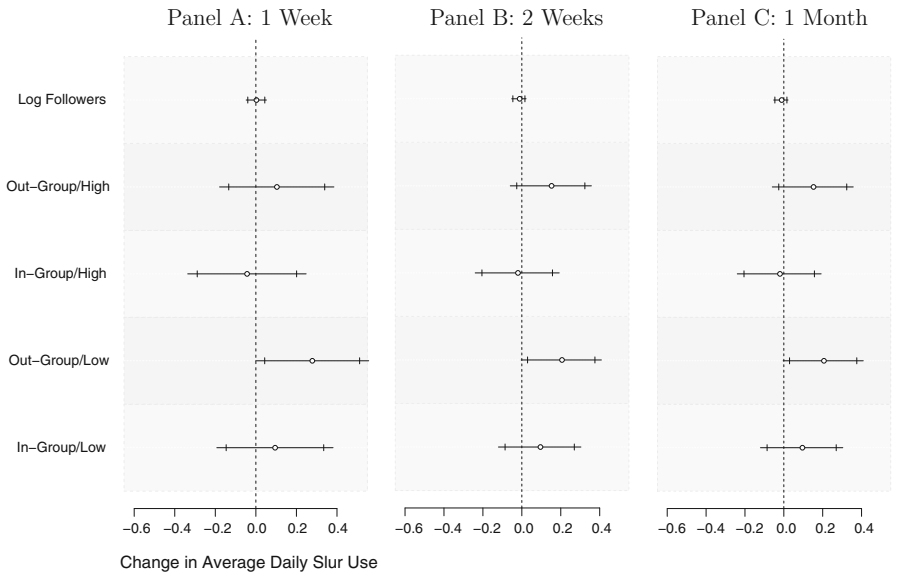

Fig. 6 Non-anonymous subjects ($N = 84$). *Each panel* represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the Out-group/Low Follower treatment in *Panel A* shows these subjects increased their average daily usage of this slur by 0.28 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the 2 months prior to the treatment. The *vertical tick marks* represent 90% confidence intervals and the *full lines* represent 95% confidence intervals

(2012) about eliciting donations to an LGTBQ organization finds that callers who self-identify as LGTBQ in an effort to personalize the issue are less effective than those who do not, and they believe that this is caused by reactance to the pressure implied by this personalization.

An example of reactance in my experiment is the subject who tweeted at my (black) bot twice: "@[bot] I DONT GIVE A FUCK N****R STFUFUCK YOU AND YOUR MOTHER" and "LMFAO N****R LOVERS NEEDA CHILL". For a subset of the subjects, reactance to the treatment actually caused a short-term increase in the use of racist language. Only around 30% of the subjects responded to the treatment, and this rate did not vary across the treatment arms.[17] Overall, this phenomenon was overwhelmed by the overall decrease in the longer time periods. Future studies should employ a larger sample size to better differentiate between these short- and long-term effects of social sanctioning.

The effect of anonymity was found to be contrary to my prediction in H_2. My expectation was that the treatment effect would be smaller for more anonymous subjects, as suggest by the findings in Omernick and Sood (2013) and Hosseinmardi et al. (2014). However, the treatment effects turned out to be smaller among less

[17] These responses also did not vary in terms of vitriol between the treatment arms. In fact, even the number of subjects that responded to call my bot a "n****r" did not vary significantly between the white and black bots.

anonymous subjects, and the treatment caused an increase in harassment for the non-anonymous subjects in the Out-group/Low Followers condition. This is consistent with the expectations of SIDE theory, and with the findings in Postmes et al. (2001).

Still, "Anonymity" in the current context does not map exactly onto these previous findings, and I urge caution in generalizing the results of this study. Specifically, these subjects selected their own level of anonymity, according to some process that is not well understood. The heterogeneous treatment effects may not represent the effects of anonymity per se, but of some other unobserved characteristic of the subjects. Future research on why people choose to remain anonymous on mixed-anonymity platforms like Twitter can help solve this puzzle.

Conclusion

Online communities represent an important development in empowering people to express themselves and communicate with the world without being limited by their physical location or social status. However, this freedom also enables some individuals to behave badly, unconstrained by social norms and uninhibited by biological feedback mechanisms restricting antisocial behavior. One manifestation of this is the harassment of members of disadvantaged groups, aiming to silence and weaken the victims of this harassment and to solidify in-group membership. In the context of the US, this often takes the form of white men harassing women and racial minorities.

To address this problem, online network administrators or government entities can explicitly ban harassing individuals or restrict certain language use. These efforts can backfire, though, and cause people to use even more racist or misogynist slurs to better differentiate themselves and their group from the "political correctness" they associate with censorship. Approaches that operate through promoting positive social norms, like the one employed in this paper, may offer a better way to develop online communities that are less toxic.

The experiment performed in this paper tests another approach to reduce the incidence of racist online harassment. By explicitly priming the subjects' membership in offline communities and updating their beliefs about the norms of online behavior, the treatment caused a significant reduction in the use of racist slurs. However, this effect was only observed among the subsample of subjects who had anonymous profiles. Among subjects who disclosed personal information, there was no significant reduction in the use of racist slurs, and there was actually an increase in the use of racist slurs among one treatment condition.

Although prejudice reduction has been widely researched, previous studies have been limited by a combination of convenience samples of undergraduate students, self-reported outcome variables, and a short measurement period that cannot measure effect persistence. Following Paluck and Green (2009)'s call for more randomized field experiments in prejudice reduction, this paper represents an improvement in all three of these dimensions: the subjects were drawn from the general population and selected because they engaged in public harassment, the

outcome variable was behavioral and objective, and the measurement period was continuous and 2 months long.

This method, of performing experiments on subjects on social media using accounts the experimenter controls, can be applied to many contexts in which the outcome of interest is online speech. An important extension to this study would be a manipulation to reduce misogynist online harassment, which continues to be a large problem for women on social media. More broadly, it could be used to experimentally determine the best method to dissuade people on social media from communicating false and potentially dangerous information about, for example, vaccinations. However, the findings from this study do not trivially generalize to offline communication or behavior.

Although this study's demonstration of a method to reduce the expression of prejudice online is valuable in and of itself, the question remains as to whether this effect changes underlying prejudiced attitudes or behavior in the physical world. Ideally, future contributions in this area of study should aim to measure all three out of these outcomes.

Acknowledgments I would like to thank ChrisDawes, Neal Beck, Eric Dickson, James Hodgdon Bisbee, DavidBroockman, Livio Di Lonardo, Ryan Enos and Drew Dimmery, along withthree anonymous reviewers; participants at the 2015 Summer MethodsMeeting, the Harvard Experimental Political Science Graduate StudentConference, Neal's Seminar, the Yale ISPS Experiments Workshop andthe NYU Graduate Political Economy Seminar; and members of the NYUSocial Media and Political Participation (SMaPP) Lab, for theirvaluable feedback on earlier versions of this project.

Compliance with Ethical Standards

Conflict of interest The author declares that he had no conflicts of interest with respect to his authorship or the publication of this article.

Ethical Standards All procedures performed in studies involving human participants were in accordance with the ethical standards of the New York University Institutional Review Board.

Appendix

Conservative Assumption for Main Results

For the subjects who produced too few post-treatment tweets to calculate an rate of racist language use, I assumed that their post-treatment rate of racist language use was zero. This assumption makes sense substantively, because these people were no longer tweeting (and thus no longer engaging in racist harassment). However, a more conservative assumption would be to assume that there was no change in their behavior, and to assign them a post-treatment rate equal to the their pre-treatment rate. This does not substantively change the results, although the magnitude of the effect sizes becomes slightly smaller.

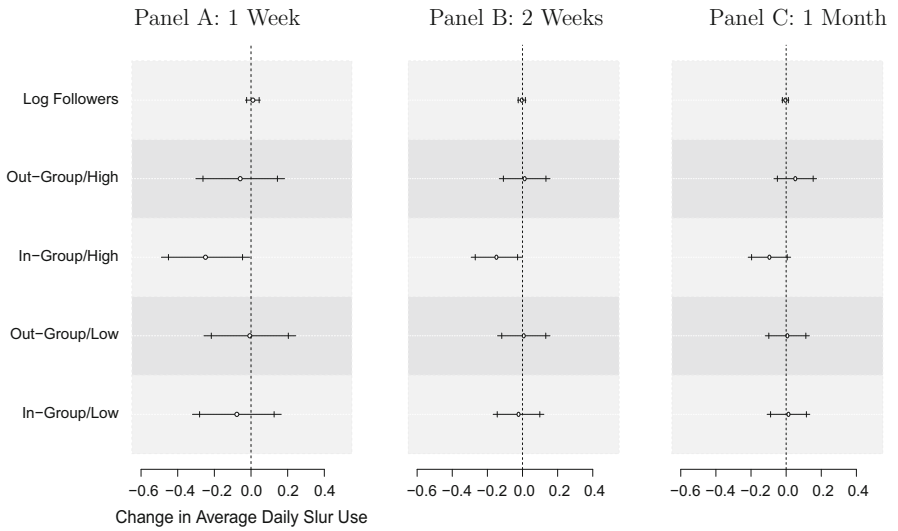

Fig. 7 Full sample ($N = 242$). *Each panel* represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "n****r" per day in the specified time period. For example, the coefficient associated with the In-group/High Followers treatment in *Panel A* shows these subjects reduced their average daily usage of this slur by 0.25 more than subjects in the control in the week after treatment. Each regression also controls for the subject's absolute daily use of this slur in the 2 months prior to the treatment. The *vertical tick marks* represent 90% confidence intervals and the *full lines* represent 95% confidence intervals

References

Allport, G. W. (1954). *The nature of prejudice*. Basic Books.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, *337*(6092), 337–341.

Banks, A. J. (2014). The public's anger: White racial attitudes and opinions toward health care reform. *Political Behavior*, *36*(3), 493–514.

Banks, A. J. (2016). Are group cues necessary? How anger makes ethnocentrism among whites a stronger predictor of racial and immigration policy opinions. *Political Behavior,* 1–23.

Bertrand, M., & Mullainathan, S. (2003). *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination*. Cambridge: National Bureau of Economic Research.

Binder, J., Zagefka, H., Brown, R., Funke, F., Kessler, T., Mummendey, A., et al. (2009). Does contact reduce prejudice or does prejudice reduce contact? A longitudinal test of the contact hypothesis among majority and minority groups in three European countries. *Journal of Personality and Social Psychology*, *96*(4), 843.

Blanchard, F. A., Crandall, C. S., Brigham, J. C., & Vaughn, L. A. (1994). Condemning and condoning racism: A social context approach to interracial settings. *Journal of Applied Psychology*, *79*(6), 993.

Bordia, P. (1997). Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *Journal of Business Communication*, *34*(1), 99–118.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues*, *55*(3), 429–444.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE pp. 71–80.

Chhibber, P., & Sekhon, J. S. (2014). The asymmetric role of religious appeals in India.
Coppock, A., Guess, A., & Ternovski, J. (2015). When treatments are tweets: A network mobilization experiment over twitter. *Political Behavior,* 1–24.
Crandall, C. S., Eshleman, A., & O'Brien, L. (2002). Social norms and the expression and suppression of prejudice: The struggle for internalization. *Journal of personality and social psychology*, *82*(3), 359.
Dovidio, J. F., & Gaertner, S. L. (1999). Reducing prejudice combating intergroup biases. *Current Directions in Psychological Science*, *8*(4), 101–105.
Gulker, J. E., Mark, A. Y., & Monteith, M. J. (2013). Confronting prejudice: The who, what, and why of confrontation effectiveness. *Social Influence*, *8*(4), 280–293.
Harrison, B. F., & Michelson, M. R. (2012). Not that theres anything wrong with that: The effect of personalized appeals on marriage equality campaigns. *Political Behavior*, *34*(2), 325–344.
Henson, B., Reyns, B. W., & Fisher, B. S. (2013). Fear of crime online? Examining the effect of risk, previous victimization, and exposure on fear of online interpersonal victimization. *Journal of Contemporary Criminal Justice*.
Hinduja, S., & Patchin, J. W. (2007). Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, *6*(3), 89–112.
Hosseinmardi, H., Rafiq, R. I., Li, S., Yang, Z., Han, R., Mishra, S., & Lv, Q. (2014). A comparison of common users across instagram and ask. fm to better understand cyberbullying. *arXiv preprint*arXiv:1408.4882 .
Kam, C. D., & Kinder, D. R. (2012). Ethnocentrism as a short-term force in the 2008 American presidential election. *American Journal of Political Science*, *56*(2), 326–340.
Kennedy, M. A., & Taylor, M. A. (2010). Online harassment and victimization of college students. *Justice Policy Journal*, *7*(1), 1–21.
Kiesler, S., Siegel, J., & McGuire, T. W. (1984). Social psychological aspects of computer-mediated communication. *American Psychologist*, *39*(10), 1123.
Lea, M., & Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies*, *34*(2), 283–301.
Mantilla, K. (2013). Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, *39*(2), 563–570.
Moor, P. J. (2007). Conforming to the flaming norm in the online commenting situation.
Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.
Omernick, E., & Sood, S. O. (2013). The impact of anonymity in online communities. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE pp. 526–535.
Paluck, E. L., & Green, D. P. (2009). Prejudice reduction: What works? A review and assessment of research and practice. *Annual review of psychology*, *60*, 339–367.
Paluck, E. L., Shepherd, H., & Aronow, P. M. (2016). Changing climates of conflict: A social network experiment in 56 schools. *Proceedings of the National Academy of Sciences*, *113*(3), 566–571.
Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*(5), 751.
Piston, S. (2010). How explicit racial prejudice hurt Obama in the 2008 election. *Political Behavior*, *32*(4), 431–451.
Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811.
Postmes, T., Spears, R., Sakhel, K., & Groot, D. D. (2001). Social influence in computer-mediated communication: The effects of anonymity on group behavior. *Personality and Social Psychology Bulletin*, *27*(10), 1243–1254.
Rasinski, H. M., & Czopp, A. M. (2010). The effect of target status on witnesses' reactions to confrontations of bias. *Basic and Applied Social Psychology*, *32*(1), 8–16.
Reicher, S. D., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198.
Shepherd, H., & Paluck, E. L. (2015). Stopping the drama gendered influence in a network field experiment. *Social Psychology Quarterly*, *78*(2), 173–193.
Sherif, M., & Sherif, C. W. (1953). Groups in harmony and tension; an integration of studies of intergroup relations.
Sood, S., Antin, J., & Churchill, E. (2012). Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM pp. 1481–1490.
Stangor, C., Sechrist, G. B., & Jost, J. T. (2001). Changing racial beliefs by providing consensus information. *Personality and Social Psychology Bulletin*, *27*(4), 486–496.

Stringhini, G., Egele, M., Kruegel, C., & Vigna, G. (2012). Poultry markets: On the underground economy of twitter followers. In *Proceedings of the 2012 ACM workshop on Workshop on online social networks*. ACM pp. 1–6.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, *33*(47), 74.

Walther, J. B. (1996). Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, *23*(1), 3–43.

Weaver, V. M. (2012). The electoral consequences of skin color: The hidden side of race in politics. *Political Behavior*, *34*(1), 159–192.

Xu, Z., & Zhu, S. (2010). Filtering offensive language in online communities using grammatical relations. *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB* 2.

Zitek, E. M., & Hebl, M. R. (2007). The role of social norm clarity in the influenced expression of prejudice over time. *Journal of Experimental Social Psychology*, *43*(6), 867–876.