Cleaning up the Online Commons: Experimentally Reducing Racist Harassment

Kevin Munger

February 4, 2016

Abstract

I conduct an experiment which examines the impact of group norm promotion and social sanctioning on racist online harassment. Racist online harassment demobilizes 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 between high and low status (by varying the number of Twitter followers each bot has), I find that subjects who were sanctioned by a high-status 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 two-month data collection period. This represents an advance in the study of prejudiced behavior.

1 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, Reyns, and Fisher, 2013), systematically de-mobilizing the populations who tend to be victimized. 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 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 racial slur 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 controled ("bots"), I tweeted at the subjects to tell them that their behavior was unacceptable. I varied two aspects of the bots, resulting in a 2x2 experimental design: the first dimension of variation was the identity of the bot, to test the theory that sanctioning by members of a person's in-group is more effective. The second variation was in the number of followers the bot has. The theory is that bots who are seen as being more influential or higher status will be more

efficacious.

I find support for the hypothesis[1] that the same message had disparate impact based on the 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. However, this effect was only found among messages sent by accounts that appeared to be higher status in the community, as measured by the 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. I also check to see if this effect was driven by substitution away from racist language into misogynistic language, and find no evidence that this occurred. However, I find only weak support for the hypothesis that these effects would larger among subjects who provided more identifying information on their profile.

Overall, in the one month post-treatment collection period, my intervention caused the 50 subjects in the most effective treatment condition to tweet the word "nigger" an estimated 186 fewer times in the month after treatment.

2 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, Eshleman, and O'Brien (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 "nigger" at another person constitutes racist harassment, regardless of how justified the user beliefs 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

[1]All hypotheses were pre-registered at EGAP.org prior to any data collection.

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 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 (Blanchard et al., 1994; Dovidio and Gaertner, 1999; Plant and Devine, 1998). 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, Sechrist, and Jost (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 racism 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 and lower comprehension (Bordia, 1997; Kiesler, Siegel, and McGuire, 1984; Walther, 1996).

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 Social Identity model of Deindividuation Effects (SIDE model), in which the depressed sense of one's personal identity is supplanted by an increased sense of one's social identity (Lea and Spears, 1991; Reicher, Spears, and Postmes, 1995).

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.

Just as prejudiced harassment against out-groups has been used to signal in-group loyalty in the physical world, 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 (Kam and Kinder, 2012; Piston, 2010).

But SIDE also suggests an avenue for reducing online racist harassment: individuals' social identity is actually composed of several overlapping online identities, and by making others of them more salient, the influence of one specific online community should be diminished.

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, Guess, and Ternovski (2015), one of the first randomized control experiments to be conducted entirely on Twitter.

The crucial advantage of this experimental design is that I can measure real behavior continuously for months. In order to quantify this behavior, I operationalize racist online harassment in the form of the use of the word "nigger." I collected this data by scraping the Twitter history of the 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 and by higher-status individuals (Gulker, Mark, and Monteith, 2013; Rasinski and Czopp, 2010). Based on these findings, my hypothesis was that the largest treatment effect would be from In-group/High status bots and that the smallest treatment effect would be from Out-group/Low status 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 status* $> \begin{smallmatrix} In-group/Low\ status \\ Out-group/High\ status \end{smallmatrix} >$ *Out-group/Low status.*

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 (Hosseinmardi et al., 2014; Omernick and Sood, 2013). 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 0 (most anonymous) to 2 (least anonymous).

Hypothesis 2 *The magnitude of the decrease in harassment will positively covary with the subject's Anonymity Score.*

3 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 (Chen et al., 2012; Yin et al., 2009). 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. The cutting-edge "Lexical Syntactic Feature-based" classifier developed by Chen et al. (2012) uses contextual information to improve on previous methods for detecting harassment; however, in corpuses that contain enough strongly offensive language, a simple Support Vector Machine using a dictionary of strongly offensive terms outperforms this more sophisticated classifier. 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 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 ("nigger") at another account.[3] In the racial context of the United States, this term is

[2]Whether an 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.

[3]As is recorded in my Pre-Analyis Plan (registered at EGAP), I had originally intended to perform

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 1,000 tweets) of anyone who tweeted the word "nigger" 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, Antin, and Churchill (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]

---

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.

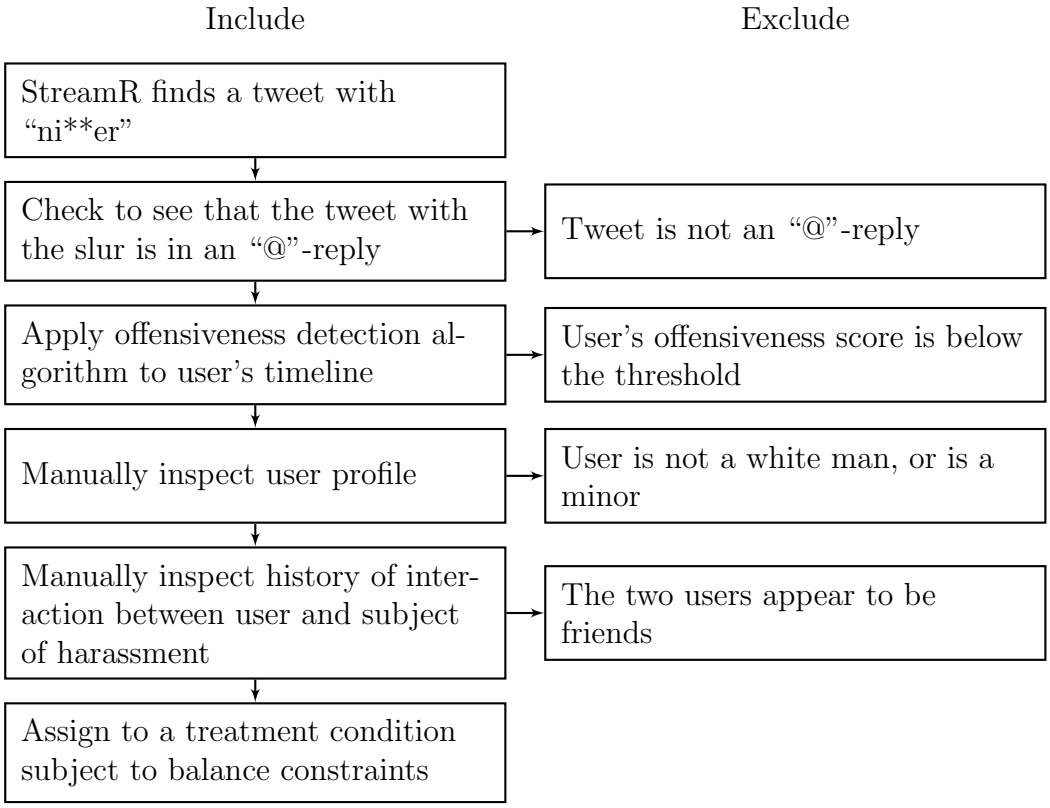
[4]Chen et al. (2012), for example, emulates Xu and Zhu (2010) and takes a list of terms from the website www.noswearing.com.

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

[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

Figure 1: Sample Selection Process



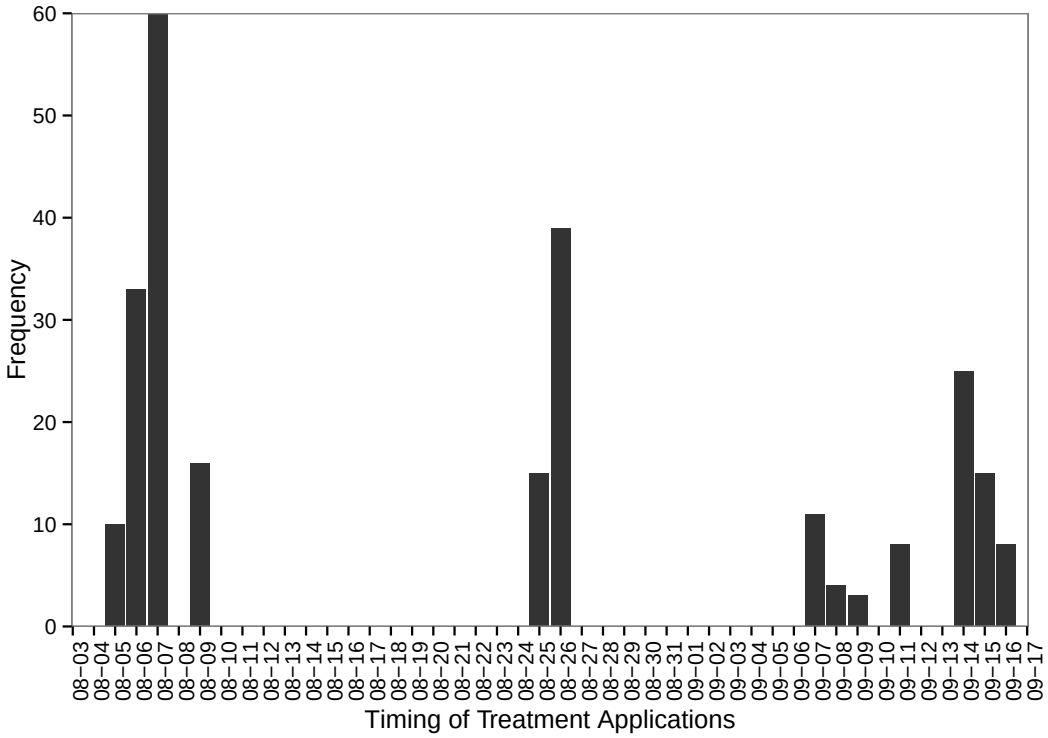
This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded.

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 online harassment of blacks, I only included subjects who were white men. This ensured that the in-groups of interest (gender and race) don't vary among the subjects, and thus that the treatments were the same. I also included anonymous users because there were a large number of such accounts engaging in prejudiced harassment. Because anonymous users might differ from less anonymous users in important ways, 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 any indication of being underage (especially mentioning high school) caused me to remove the user from the sample.

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.

the users with a long back-and-forth, with slurs interspersed with more obviously friendly terms.

Figure 2: Timing of the Experiment in the Field



The number of subjects added to the sample each day is plotted on the y-axis.

Table 1: Experimental Design and Hypothesized Effect Sizes

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

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 are 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; and the third wave from September 7th to September 11th; and the last was from September 14th to September 16th. See Figure 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 are
hurt 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 supported by findings in Gulker, Mark, and Monteith (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

[8]Throughout the assignment process, I matched subjects in each treatment group on their (0 to 2) Anonymity Score, determined by whether they provided a real name and/or a picture of themselves. They were otherwise randomly assigned.

[9]This process was approved by NYU's Internal Review Board. Note that these subjects had not given their informed consent to participate in this experiment, but that the intervention I applied falls within the "normal expectations" of their user experience on Twitter. Note also that 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.

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 varied that quantity to test the finding in Rasinski and Czopp (2010) that sanctioning by high-status individuals is more effective than that by low-status individuals.

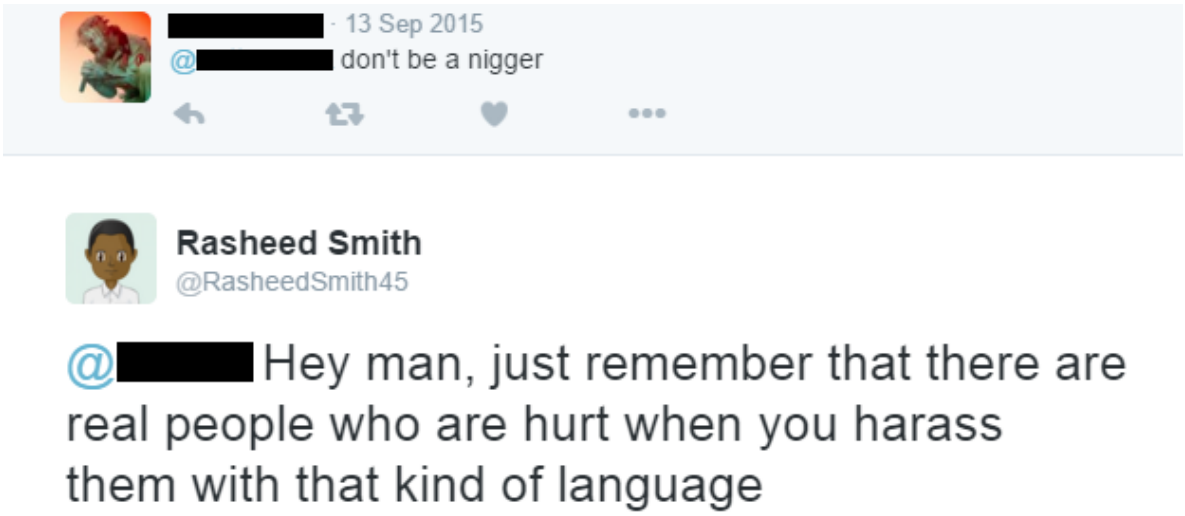
For example, users assigned to the Out-group/Low status condition were sent a message like the one seen in Figure 3(a), sent by bot @RasheedSmith45. After the subject received the treatment, he got a "notification" from Twitter, which caused him to be exposed to the treatment tweet. Because being admonished 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, Figure 3(b). @GregJackson730 was a bot in the In-group/Low Status 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 status treatment downward.

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

Figure 3: Treatments



(a) The treatment

(b) The bot applying the treatment

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).[10]

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. On possible source of skepticism was that he 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 Figure 3(b), there needed to be a 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.

4 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 two 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 231 in the sample,[11] but I

[10]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.

[11]Initially, I assigned 243 subjects to one of the 4 treatment arms or to the control group. Due to technical issues with my code for scraping Twitter, I only successfully collected pre-treatment tweets for 231 of these. However, these missing observations occurred at random, so this does not bias my results.

Table 2: Attrition Rates

	Control	A	B	C	D
Baseline # of subjects	40	49	44	50	48
# of subjects with > 1 Post-treatment tweets	40	46	42	47	47
# of subjects with > 25 Post-treatment tweets	40	34	33	35	43
Attrition %, < 25 Post-treatment tweets	10%	18%	16%	18%	4%

The number of subjects who tweeted more than 1 or 25 times after the application of the treatment.

could not distinguish between the last two.[12] Table 2 presents the attrition rates among the different treatment arms in the sample. Among the four treatment conditions, 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 14%, compared to 10% among the control subjects, an insignificant difference ($p = .44$).

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 have missing data. I made a conservative assumption: for each of these observations with missing post-treatment data, I assumed no change in their rate of racist language use pre- and post-treatment.[13]

The results support H_1. In Figure 4, Panel A shows the effect of the different treatment arms on the absolute daily use of the word "nigger" over the week after the treatment. Panel B expands the time period to two weeks, and Panel C expands it to one 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' Anonymity Score and log number of followers, displayed in the first and second rows. Each regression also controls for the average rate of the subjects' use of that offensive term in the two months prior to the treatment. The four treatment arms each represent the comparison between that arm and the control group, and each treatment

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

[13]A less conservative but more substantively plausible assumption is to treat these observations as having a post-treatment rate of offensive language use of zero–the subjects who are no longer tweeting publicly have ceased to engage in online harassment. Appendix B presents the results with this alternate assumption. The results are substantively similar, though of larger magnitude.

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 status 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 $p < .10$, and it is of smaller magnitude than the reduction in the In-group/High status condition. This was contrary to my expectations in H_1: I predicted that both the Out-group/High status and In-group/Low status conditions would have a larger effect than the Out-group/Low status condition.

Comparing across the panels of Figure 4 shows the decay in the effect of Treatment C over time. Although the effect remains statistically significant, the coefficient decreases steadily. In Panel A, the point estimate of -.27 indicates that the daily rate of the use of the word "nigger" decreased by .27 more among subjects in Treatment condition C than among subjects in the control condition. This average treatment effect for condition C decreased in magnitude to -.17 in Panel B and -.11 in Panel C. I collected data for 2 months, but these results are not shown because none of the treatments are significant.

There was, however, little support for H_2. I predicted that subjects who provided either their real name or a picture of themselves on their account (and thus had higher Anonymity Scores) would experience a greater decrease in racist language after treatment. In Panel A, the effect is in the expected direction, but has a p-value of .14. Comparing across panels A, B and C shows that this effect decreases in magnitude to almost zero, so to the extent that the effect in the one week time period is suggestive of support for H_2, this ceases to be true in longer time periods.

One risk with interventions like the one performed in this experiment is substitution: that a reduction in a targeted behavior could be offset by a increase in another behavior. In addition to racist harassment, misogynist harassment is one of the biggest problems in online communities. To check for a possible substitution effect, Figure 5 presents the same analysis as in Figure 4, except that rather than the use of a racist slur, it takes as dependent variable the use of misogynist slurs.[14] Panel A estimates the effect of all four treatment arms on the use of misogynist slurs over a one week time frame, but none of these estimate are statistically significant. The point estimates of all treatment arms become close to zero in longer time frames. This is consistent with the lack of a substitution effect.

[14]The list I used consisted of the words bitch, cunt, dyke, skank, slut, whore, and ho.

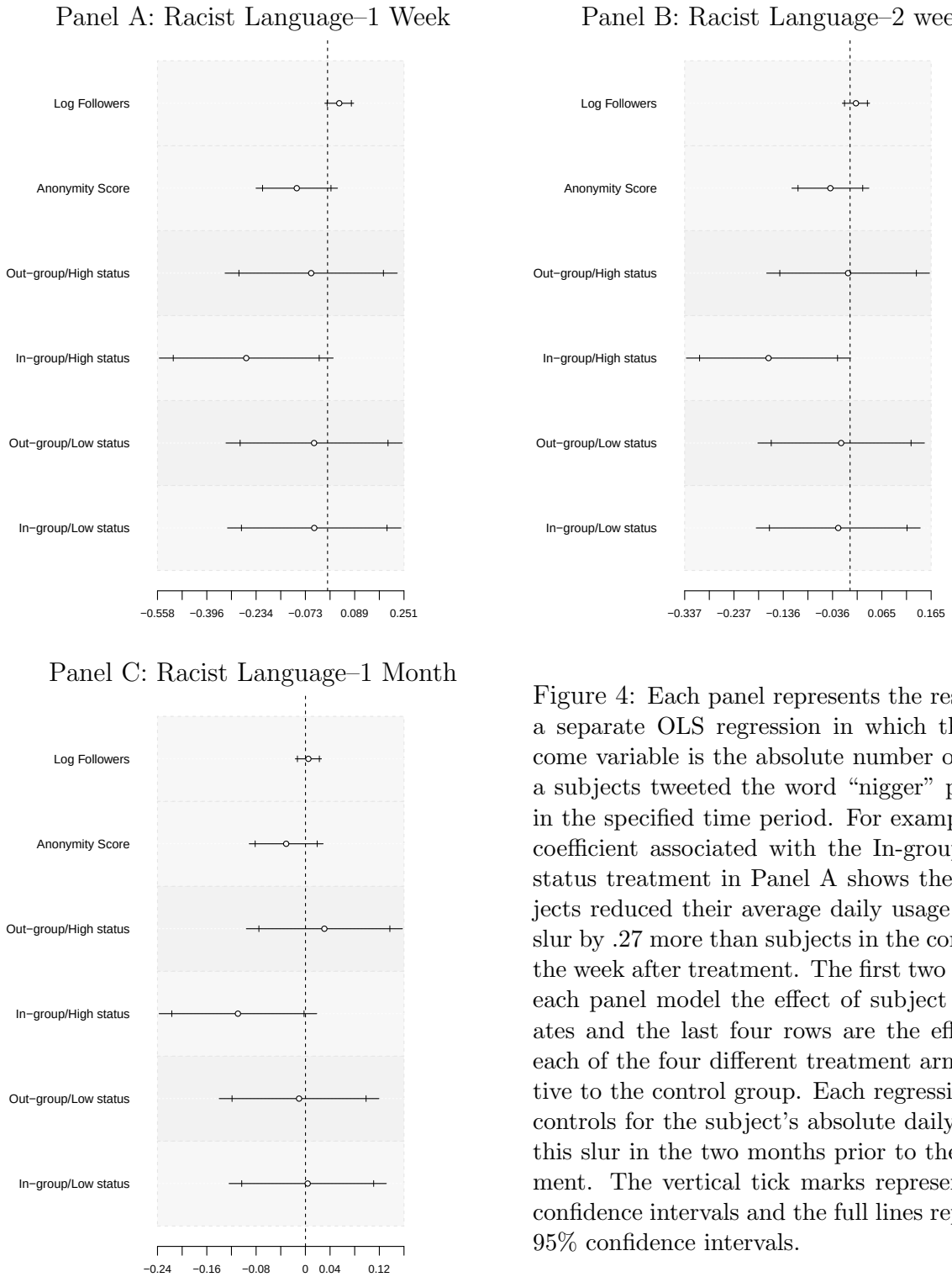


Figure 4: Each panel represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted the word "nigger" per day in the specified time period. For example, the coefficient associated with the In-group/High status treatment in Panel A shows these subjects reduced their average daily usage of this slur by .27 more than subjects in the control in the week after treatment. The first two rows of each panel model the effect of subject covariates and the last four rows are the effects of each of the four different treatment arms relative to the control group. Each regression also controls for the subject's absolute daily use of this slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

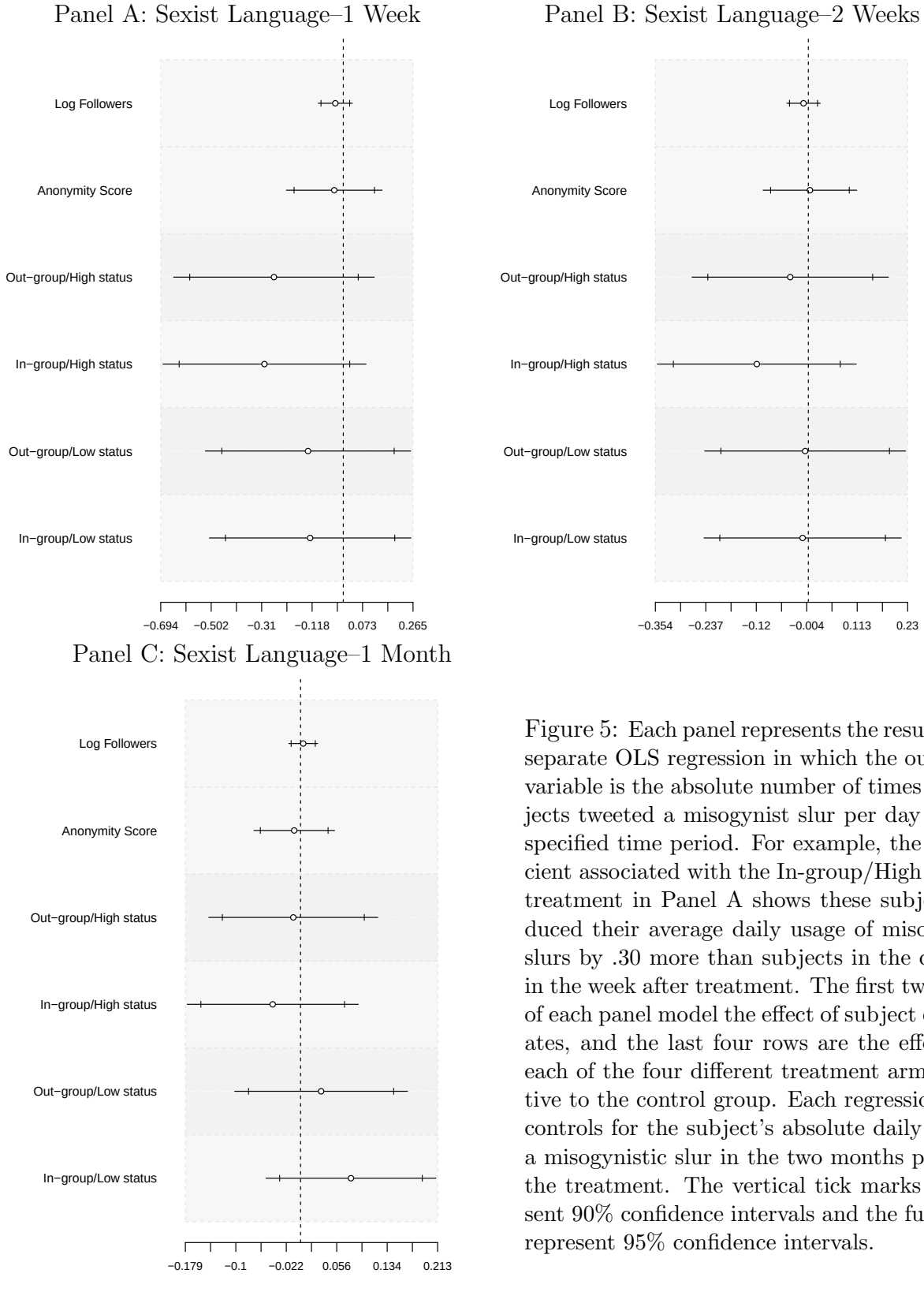


Figure 5: Each panel represents the results of a separate OLS regression in which the outcome variable is the absolute number of times a subjects tweeted a misogynist slur per day in the specified time period. For example, the coefficient associated with the In-group/High status treatment in Panel A shows these subject reduced their average daily usage of misogynist slurs by .30 more than subjects in the control in the week after treatment. The first two rows of each panel model the effect of subject covariates, and the last four rows are the effects of each of the four different treatment arms relative to the control group. Each regression also controls for the subject's absolute daily use of a misogynistic slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

5 Discussion

The primary prediction expressed in H_1, that the In-group/High Status treatment would cause the largest reduction in racist language use, was borne out. This effect was larger than either the In-group/Low status or a Out-group/ High status treatments, although these latter two reductions were not significant as expected. This prediction followed from the SIDE model: the treatment primed subjects' membership in social groups other than the racist online communities that are salient when they engage in online harassment. The treatment also caused them to update their beliefs about norms of online behavior. Encouragingly, these effects persisted over time, for the first month under study, although not for two months. Also, the effect was significant at $p < .05$ in the two week time period, but it was only significant at $p < .10$ in the one week and one 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 one 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 (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 [slur] STFUFUCK YOU AND YOUR MOTHER" and "LMFAO [slur] 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. However, 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 on the use of racist language only weakly conformed to my prediction in H_2. My expectation was that the treatment effect would be smaller for more anonymous subjects, as the theory suggests that more anonymity is associated

with higher levels on online harassment (Hosseinmardi et al., 2014; Omernick and Sood, 2013).

This was not borne out in my data. I am not confident in how to reconcile this with previous findings. Earlier studies have examined system-wide levels of anonymity, comparing rates of harassment on a forum before and after it switched from being anonymous to forcing users to identify themselves, and I may have been mistaken in applying this finding to individuals who can choose their own level of anonymity on a forum like Twitter.

I was concerned about this anti-racist treatment causing a possible substitution from racist language use to the use of misogynist language, but find no evidence that this happened.

6 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 confuse the use of racist or misogynist slurs with a defense of free speech. In the course of finding subjects and applying the treatment, I encountered two common manifestations of this phenomenon: men affiliated with "GamerGate," an online movement objecting to what they see as a progressive bias in the media, some supporters of which have harassed female journalists; and supporters of Donald Trump's ethnocentric, anti-immigration Presidential campaign.

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. This reduction was

not the result of substitution into the use of more misogynist slurs.

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 are drawn from the general population and selected because they engaged in public harassment, the outcome variable is behavioral and objective, and the measurement period is continuous and two months long.

This method, of performing experiments on subjects on social media using accounts the experimenter controls and manipulate, 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.

Although this study's evidence 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.

Appendix

A Dictionary of Offensive Terms

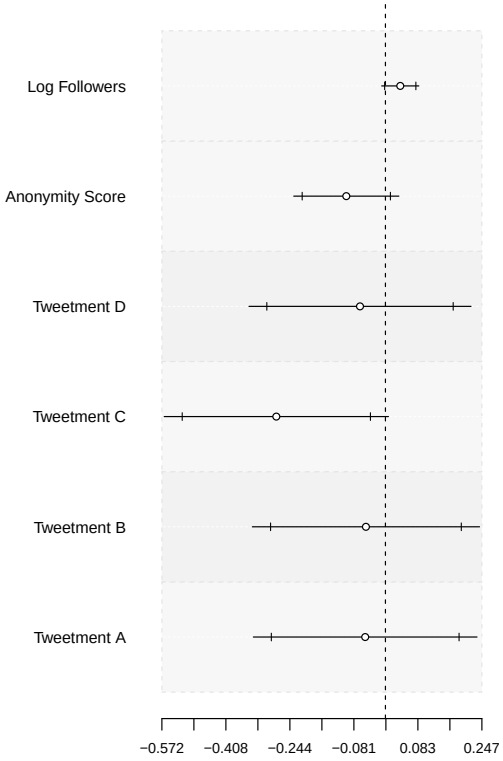
Below are the terms used to calculate the baseline level of offensiveness for users prior to including them in the sample. The "*" symbol indicates the use of regular expression, so that a word containing that exact string of characters would also count as an offensive term. For example, "shithead" contains the string "shit" and was counted by "*shit*". "japan" contains the string "jap", but was not counted because of the lack of the "*" character.

"beaner","chinc", "chink","coon", "dego", "gook", "guido", "heeb", "kike", "kyke", "jigaboo", "mick", "negro","nigger*", "niglet", "paki", "porchmonkey", "pollock", "ruski", "sandnigger", "spic", "wop", "jap", "junglebunny", "spick", "wetback" "bitch", "cunt", "dyke", "skank*", "slut*", "whore*", "ho" "*bastard*", "*shit*", "*fuck*", "carpetmuncher", "*cock*", "cum", "*douche*", "*fag*", "fudgepacker", "blowjob", "handjob", "homo", "jizz", "lesbo", "lezzie", "pussy", "queerbait", "rimjob", "skeet", "tard"

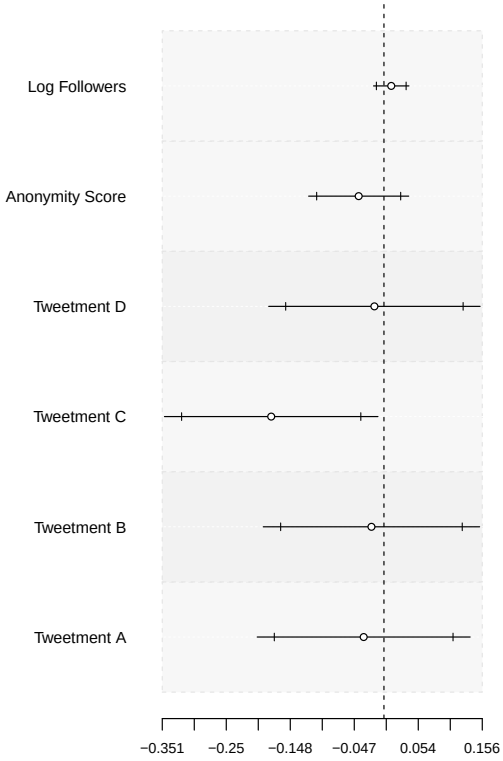
B Anti-Conservative Assumption for Main Results

For the subjects who produced too few post-treatment tweets to calculate an rate of racist language use, I assumed no change in their behavior pre- and post-treatment. Because these people were no longer tweeting (and thus no longer engaging in racist harassment), however, it makes more sense substantively to assume for these observations a rate of racist language use equal to zero. I replicate the main analysis below, except with this anti-conservative assumption. This does not substantively change the results, although the magnitude of the effect sizes becomes larger.

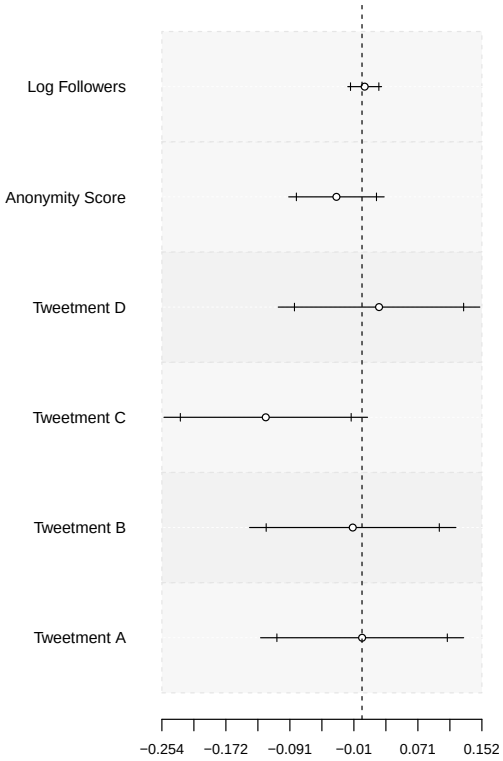
Panel A: Racist Language–1 Week



Panel B: Racist Language–2 weeks



Panel C: Racist Language–1 Month



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 "nigger" per day in the specified time period. For example, the coefficient associated with the In-group/High status treatment in Panel A shows these subjects reduced their average daily usage of this slur by .29 more than subjects in the control in the week after treatment. The first two rows of each panel model the effect of subject covariates and the last four rows are the effects of each of the four different treatment arms relative to the control group. Each regression also controls for the subject's absolute daily use of this slur in the two months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.

References

Allport, Gordon Willard. 1954. *The nature of prejudice*. Basic books.

Banks, Antoine J. 2014. "The publics anger: White racial attitudes and opinions toward health care reform." *Political Behavior* 36 (3): 493–514.

Banks, Antoine J. 2016. "Are Group Cues Necessary? How Anger Makes Ethnocentrism Among Whites a Stronger Predictor of Racial and Immigration Policy Opinions." *Political Behavior* pp. 1–23.

Bertrand, Marianne, and Sendhil Mullainathan. 2003. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. Technical report National Bureau of Economic Research.

Binder, Jens, Hanna Zagefka, Rupert Brown, Friedrich Funke, Thomas Kessler, Amelie Mummendey, Annemie Maquil, Stephanie Demoulin, and Jacques-Philippe Leyens. 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, Fletcher A, Christian S Crandall, John C Brigham, and Leigh Ann Vaughn. 1994. "Condemning and condoning racism: A social context approach to interracial settings." *Journal of Applied Psychology* 79 (6): 993.

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

Brewer, Marilynn B. 1999. "The psychology of prejudice: Ingroup love and outgroup hate?" *Journal of social issues* 55 (3): 429–444.

Chen, Ying, Yilu Zhou, Sencun Zhu, and Heng Xu. 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, Pradeep, and Jasjeet S Sekhon. 2014. "The asymmetric role of religious appeals in India.".

Coppock, Alexander, Andrew Guess, and John Ternovski. 2015. "When Treatments are Tweets: A Network Mobilization Experiment over Twitter." *Political Behavior* pp. 1–24.

Crandall, Christian S, Amy Eshleman, and Laurie O'Brien. 2002. "Social norms and the expression and suppression of prejudice: the struggle for internalization." *Journal of personality and social psychology* 82 (3): 359.

Dovidio, John F, and Samuel L Gaertner. 1999. "Reducing prejudice combating intergroup biases." *Current Directions in Psychological Science* 8 (4): 101–105.

Gulker, Jill E, Aimee Y Mark, and Margo J Monteith. 2013. "Confronting prejudice: The who, what, and why of confrontation effectiveness." *Social Influence* 8 (4): 280–293.

Harrison, Brian F, and Melissa R Michelson. 2012. "Not that theres anything wrong with that: The effect of personalized appeals on marriage equality campaigns." *Political Behavior* 34 (2): 325–344.

Henson, Billy, Bradford W Reyns, and Bonnie S Fisher. 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* p. 1043986213507403.

Hosseinmardi, Homa, Rahat Ibn Rafiq, Shaosong Li, Zhili Yang, Richard Han, Shivakant Mishra, and Qin Lv. 2014. "A Comparison of Common Users across Instagram and Ask. fm to Better Understand Cyberbullying." *arXiv preprint arXiv:1408.4882* .

Kam, Cindy D, and Donald R Kinder. 2012. "Ethnocentrism as a short-term force in the 2008 American presidential election." *American Journal of Political Science* 56 (2): 326–340.

Kennedy, M Alexis, and Melanie A Taylor. 2010. "Online harassment and victimization of college students." *Justice Policy Journal* 7 (1).

Kiesler, Sara, Jane Siegel, and Timothy W McGuire. 1984. "Social psychological aspects of computer-mediated communication." *American psychologist* 39 (10): 1123.

Lea, Martin, and Russell Spears. 1991. "Computer-mediated communication, de-individuation and group decision-making." *International Journal of Man-Machine Studies* 34 (2): 283–301.

Mantilla, Karla. 2013. "Gendertrolling: Misogyny Adapts to New Media." *Feminist Studies* pp. 563–570.

Moor, Peter J. 2007. "Conforming to the flaming norm in the online commenting situation.".

Nyhan, Brendan, and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior* 32 (2): 303–330.

Omernick, Eli, and Sara Owsley Sood. 2013. The Impact of Anonymity in Online Communities. In *Social Computing (SocialCom), 2013 International Conference on.* IEEE pp. 526–535.

Paluck, Elizabeth Levy, and Donald P Green. 2009. "Prejudice reduction: What works? A review and assessment of research and practice." *Annual review of psychology* 60: 339–367.

Pettigrew, Thomas F, and Linda R Tropp. 2006. "A meta-analytic test of intergroup contact theory." *Journal of personality and social psychology* 90 (5): 751.

Piston, Spencer. 2010. "How explicit racial prejudice hurt Obama in the 2008 election." *Political Behavior* 32 (4): 431–451.

Plant, E Ashby, and Patricia G Devine. 1998. "Internal and external motivation to respond without prejudice." *Journal of Personality and Social Psychology* 75 (3): 811.

Rasinski, Heather M, and Alexander M Czopp. 2010. "The effect of target status on witnesses' reactions to confrontations of bias." *Basic and Applied Social Psychology* 32 (1): 8–16.

Reicher, Stephen D, Russell Spears, and Tom Postmes. 1995. "A social identity model of deindividuation phenomena." *European review of social psychology* 6 (1): 161–198.

Sherif, Muzafer, and Carolyn W Sherif. 1953. "Groups in harmony and tension; an integration of studies of intergroup relations.".

Sood, Sara, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM pp. 1481–1490.

Stangor, Charles, Gretchen B Sechrist, and John T Jost. 2001. "Changing racial beliefs by providing consensus information." *Personality and Social Psychology Bulletin* 27 (4): 486–496.

Stringhini, Gianluca, Manuel Egele, Christopher Kruegel, and Giovanni Vigna. 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.

Walther, Joseph B. 1996. "Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction." *Communication research* 23 (1): 3–43.

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

Xu, Zhi, 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*.

Yin, Dawei, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. "Detection of harassment on web 2.0." *Proceedings of the Content Analysis in the WEB* 2.

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