Measuring Tweetment Effects: Social Norm Promotion on Online Harassers

Kevin Munger

October 2, 2015

Abstract

I conduct an experiment which examines the impact of group norm promotion and social sanctioning on harassment and other manifestations of prejudice. I collect a sample of Twitter users who have harassed other users and use accounts I control to sanction the harassing behavior. By varying the identity of the bots that apply social sanctioning, I test theories about social norm promotion by in-group and out-group individuals. Novel contributions to the literature on prejudice reduction include real-world measures of prejudice and the ability to continuously measure treatment effects over an indefinite time horizon.

1 Introduction

The explicit expression of prejudice is no longer acceptable in mainstream US society. This is evidence for changing social norms, though implicit prejudice may continue to be a major problem. Despite this progress, the rise of online social interaction has brought with it new opportunities for some individuals to express their prejudices and engage in other forms of verbal harassment.

Much of the research on the phenomenon of online harassment or "cyber-bullying" or "flaming" has focused on its impact on children and adolescents (Dinakar, Reichart, and Lieberman, 2011; Vandebosch and Van Cleemput, 2009; Ybarra et al., 2012), but

it is an issue of serious concern for adult victims of prejudice as well. Women, minorities, and other vulnerable populations are frequently the subject of online harassment on social media sites, often in response to vocalizing views that harassers disagree with (Kennedy and Taylor, 2010; Mantilla, 2013). 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. More common is the use of profanity or slurs, personal insults and off-topic posts (spam); while not as serious as threats of violence, these forms of harassment can cause psychological harm to their victims and make them less likely to use social media in the future. Harassment of all degrees of seriousness should be divided into two categories: *prejudiced harassment*, which targets individuals because of their race, religion, gender, or other personal attribute; and *general harassment*, which involves non-specific offensiveness and targets individuals because of some kind of divergent views on, for example, sports or politics. Though this paper is about *prejudiced harassment*, both types contribute to a toxic online community.

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. The individual-level effects of these efforts on the propensity for harassment are difficult to measure. 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 well-identified, and calls for more experimental research. I conduct a novel experiment that engages with the experimental social psychological literature on prejudice reduction that is able to measure the effect of specific interventions on the real-world harassing behavior of Twitter users.

Using the "@"-structure of Twitter and the presence of profanity to identify users (with public Twitter accounts) who may be harassing other Twitter users, I collected the most recent 1,000 tweets from each of these potential subjects to establish a baseline rate of offensiveness. The experiment tests psychological findings on social sanctioning and norm promotion by having Twitter accounts that I control ("bots") tweet at the harasser that their behavior is unacceptable. There are two aspects of the bots that I vary, resulting in a 2x2 experimental design: the first dimension of variation is the identity of the bot doing the sanctioning, to test the theory that sanctioning by members of a person's in-group is more effective. The second variation is 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. The subjects of the experiment are those who have engaged in racist online harassment, and the variation in the identity of the bots is between white and black men.

I find support for the hypothesis that the same message has disparate impact based on the in-group identity (race), with messages sent by white men causing the largest reduction in offensive behavior among a subject pool of white men. Tthere is no evidence that people respond more strongly to accounts that are higher status in the community, as measured by the number of Twitter followers.

2 Literature Review

Using Yin et al. (2009)'s definition of online harassment as "deliberate annoyance," online *prejudiced harassment* is taken to mean deliberate annoyance directed at an individual because of their personal characteristics.

Prejudiced harassment presupposes the existence 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 summarizes the debate as to whether prejudice is based on necessarily inaccurate or unfounded belief; I use Crandall, Eshleman, and O'Brien (2002)'s more agnostic definition.

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. Binder et al. (2009) state the problem succinctly: "Does Contact Reduce Prejudice or Does Prejudice Reduce Contact?", and they find that the answer is yes to both. The most comprehensive review to date finds only mild support for the contact hypothesis (Pettigrew and Tropp, 2006).

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 situation" Sherif and Sherif (1953). Attitudes towards outgroups are a particularly important set of group

norms, and prejudice towards outgroups can be a strong signal of ingroup membership (Brewer, 1999).

Recent experiments have set 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 flaw common to pyschology experiments run with convenience samples: they cannot track either long-term or real-world manifestations of prejudice. Two exceptions to the former problem are Stangor, Sechrist, and Jost (2001), who show that providing consensus information about ingroup 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 or black people) up to a month after the experiment. By studying the behavior of people on Twitter, my approach is able to capture a more fine-grained measure of prejudice reduction over time.

Manifestations of prejudice online are similar to those in real life, but there are important and systematic differences. Online behavior falls under the umbrella of Comupter Mediated Communication (CMC), which has been shown to differ from face to face communication: CMC tends to result in more self-disclosure, produce more ideas and encourage more evenly distributed participation, but also 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 when using 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. Early theories of deindividuation posited that individuals' reduction of self brings more fundamental, universal human feelings to the surface (Postmes and Spears, 1998; Roeckelein, 1998). Recent research favors 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). For more anomic individuals (those lacking clear social norms to which to conform), these theories are not necessarily distinct.

However, online communities make this kind of anomie much less common, and SIDE predicts that increased anonymity makes any kind of group identification stronger.

On non-anonymous online forum like Facebook, for example, the communities are more of an extension of real-world communities, and Facebook is characterized less by online-only communities than is a purely anonymous online forum like 4chan or Reddit. 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. This can explain why case studies of of online harassment in different communities have found its prevalence to vary from high (Alonzo and Aiken, 2004) to low (Coleman, Paternite, and Sherman, 1999).

Just as prejudiced harassment against outgroups has been used to signal ingroup loyalty in the real world, it serves the same purpose in online communities. Indeed, SIDE predicts that prejudice should be even more widespread online. But SIDE also suggests an avenue for reducing online prejudiced 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 lacunua. It also represents, with Coppock, Guess, and Ternovski (2015), one of the first randomized control experiments to be conducted entirely on Twitter.

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 prejudiced harassment can be measured, it is essential to have a sample of users who engage in prejudiced 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 *prejudiced* online harassment, I created a sample of individuals who tweeted a racial slur ("ni**er") at another account.[1] 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 not doubt that these people are aware that directly this term at another person constitutes harassment.

Using the streamR package for R, I set up a scraper to get the user information (including the most recent 1,000 tweets) of anyone who tweeted the word "ni**er" at another user. For each of these users, I applied a simple dictionary method to calculate the ratio of offensive to non-offensive words in the text of those tweets to generate an offensiveness score for that user. Sood, Antin, and Churchill (2012) warns against taking off-the-shelf dictionaries of offensive terms and importing them to a distinctive context like Twitter, and in many cases this warning should have been heeded.[2]

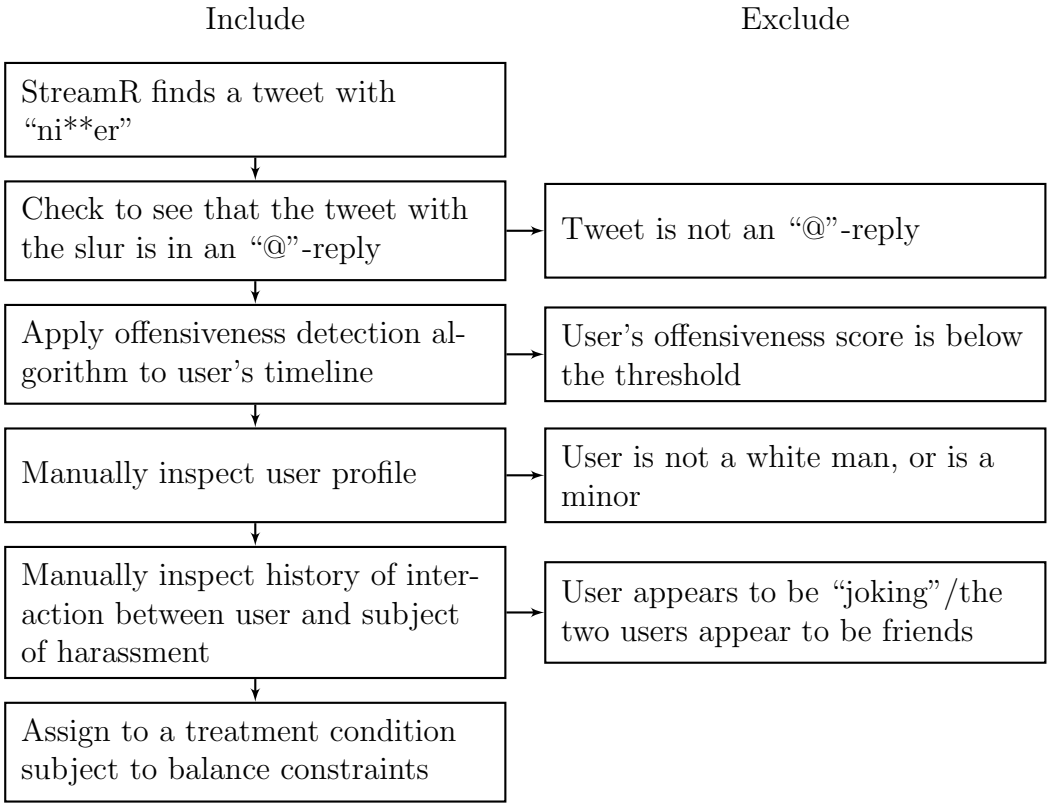
I then discarded users whose offensiveness score fell below a certain threshold and who were thus not regularly offensive. To determine what this "regularly offensive" threshold should be, I used the a random-number generator to randomly sample 450 Twitter users whose accounts were at least 6 months old.[3] I calculated the offensiveness score for these users' most recent 400 tweets and set the threshold for inclusion in the experimental sample at the 90th percentile of offensiveness.

[1] As is recorded in my Pre-Analyis Plan (registered at EGAP), I had originally inteded to preform two similar experiments: one on racist harassment, and the other on sexist harassment. However, my method was insufficent for generating a large enough sample of sexist users. For any sexist 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 sexist harassment experiment is unfeasible.

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

[3] 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.

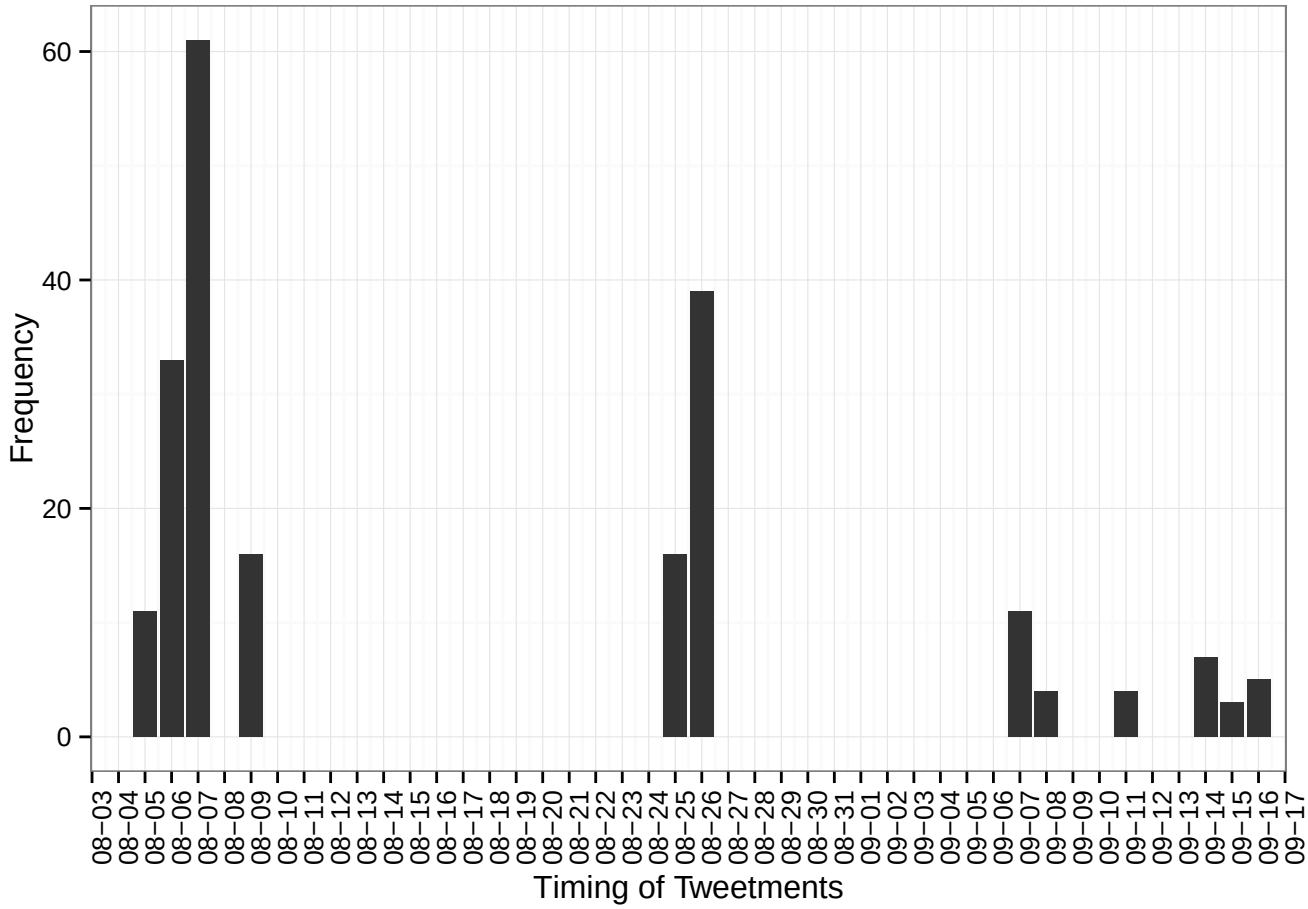
Figure 1: Sample Selection Process



This addresses many problems that 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 misclassification less likely. 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 "joking."

There were several other restrictions I placed on the sample of users. I only included subjects who were identifiable as white men or whose race and gender were unidentifiable. This was to ensure that the in-groups of interest (gender and race) don't vary among the users, and thus that the treatments are the same. I also found white men to be far and away the most common demographic engaging in prejudiced online harassment, which makes them the most representative demographic upon which to perform the experiment. I also excluded minors from the sample. Most users do not provide their exact age, but any indication of being underage (especially being in high school)

Figure 2: Timing of the Experiment in the Field



caused a user to be removed from the sample.

Because the subjects in this experiment are 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.[4] Because this process was time-consuming, and there are a fixed number of potential subjects meeting these criteria tweeting, the subject discovery and vetting took place in several periods. The first wave of subjects were collected from August 5th to August 7th, 2015; the second wave from August 25th to August 26th; and the third wave from

[4]Throughout the assignment process, I matched subjects in each treatment group on their (0 to 2) Anonymity Score. They were otherwise randomly assigned.

Table 1: Experimental Design: Racist Slurs

	In-group	Out-group
Low followers	A	B
High followers	C	D

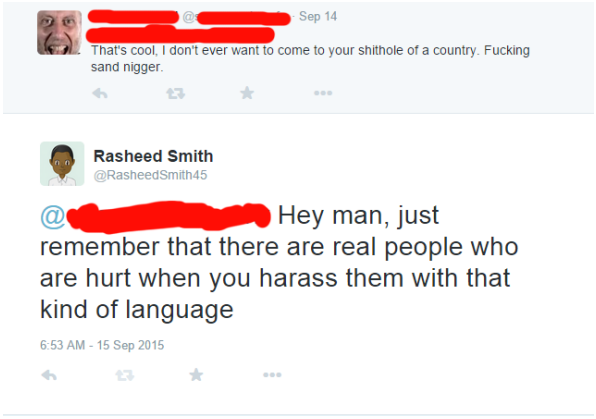
September 7th to September 11th; and the last was from September 14th to September 16th. See Figure 2 for a visual summary.[5] The crucial advantage of this real-time detection is the time that elapsed between when a user tweeted a slur and when he received the treatment is under 24 hours, adding to 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 appropriate bots (I had a total of 11) in that condition to tweet 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 is an "@"-reply, it will only show up in the Twitter feed of people who follow both of us, though it is visible to anyone who clicks on the tweet that to which my bot is replying. The four experimental conditions are summarized in Table 1. For example, users assigned to condition B (out-group, low followers) would be sent a message like the one seen in Figure 3(a), sent by @RasheedSmith45, who also has few followers. After the subject receives the treatment, he gets a "notification" from Twitter, which causes him to the tweet that constitutes the treatment. Because being admonished by a stranger is an uncommon (though far from unknown) experience, the subject will be inclined to click on the bots' account; if he does, he will see the bot's profile page, Figure 3(b); @GregJackson730 is a bot in condition A (in-group, low followers). This will all the subject to clearly determine the race and gender of his admonisher, as well as see how many followers the account has (in this case, 1). I could not, however, directly measure this behavior, and it is possible that a substantial portion of subjects did not click on the bot's profile. If that is the case, they would still have picked up the bot's race, but they would not have seen the number of followers.

As the two bots shown in Figure 3 illustrate, the variation in the bot identity will be accomplished by changing the number of followers, profile picture, username, and full

[5]This process was approved by NYU's Internal Revue Board. Note that these subjects have not given their informed consent to participate in this experiment, but that the intervention I apply falls within the "normal expectations" of their user experience on Twitter. Note also that the subjects were not debriefed. The benefits to their debriefing were deemed not to outweigh the risks to me, the researcher, in providing my personal information to a group of people with a demonstrated propensity for online harassment.

Figure 3: Treatments



(a) The treatment

(b) The bot applying the treatment

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 have 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 have between 500 and 550 followers.

When generating the bots, I chose handles that consist of first and last names that are identifiably male, female, 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 are 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 is their profile picture. It is the first thing the subject sees, and is 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, as the practice of 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).

In order to ensure that the actual treatment experienced by the subject is as similar to the "real life" experience of being sanctioned by a stranger on Twitter, it is essential that the subject be unaware that my bot is in fact a bot. If the subject suspects that the bot is not the authentic online manifestation of a concerned citizen, the effects

of norm promotion will be attenuated and the measured treatment effect will be a conservative estimate of the true treatment effect. This suspicion could arise from several possible(though unlikely) sources. The followers I bought were not high-quality followers, in that they were obviously not real accounts, but having fake or "spam" accounts is not uncommon. If the subject does a reverse image search for the profile picture of the bot, they will find that the image is of a politician with a name that does not correspond to the bot's name.

The history of tweets by the bot represents the most serious problem for verisimilitude. Under the "Tweets" tab displayed in Figure 3(b), there needs to be a plausible history of tweets to convey that this is 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 non-political news articles. However, in the default profile display, tweets that are directed "@" another user are not visible. If the subject clicks on the "Tweets & replies" tab, they become visible, but my innocuous tweets are interspersed so that the treatment tweets represent less than half of the bot's overall tweets.

I added subjects to the 4 treatment groups and 1 control group until I had a total of 245 subjects, the number I registered in my Pre-Analysis Plan. For covariate balance information, see Appendix XXX.

4 Hypotheses

Because the experiment took place under "real-world" conditions, there were a number of different ways the subjects could respond to the treatment, all of which were recorded. The subject could respond to the treat by tweeting back at the bot, or they could retweet the bot's sanctioning tweet. The subject might delete the offensive tweet, or go back and delete previous tweets.[6] The subject might block the bot. Finally, the subject might make their account private, so that no one can view it.

These behaviors were uncommon, so there are two main Dependent Variables: the overall offensiveness level and the frequency of prejudiced online harassment in the form of the use of racial slurs. These were calculated based on the Twitter history of the subject before and after being treated. I had no differential hypothesis as to which

[6]Twitter's API specifies that collections of tweets need to be updated to reflect deletions. I am thus forced to delete the content of any deleted tweets, but I can count how many there are.

of these measures of harassment will vary more after treatment; I refer to them as "harassment" in this section, and explain how they differ below.

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). My hypothesis for the relative impact of the treatments is that the effects of the bots being in-group and having more followers will be additive, of roughly the same magnitude, and that the weakest treatment will not be distinguishable from no treatment.

Hypothesis 1 *The ranking of the magnitudes of the decrease in harassment will be:* $C > A = D > B = Control.$

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[7] and whether a given name was present in their username or handle. I used these to create a categorical Anonymity Score that ranges 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.*

5 Results

The primary outcome of interest is the change in the subjects' levels of offensiveness in the four different treatment arms.

References

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

[7]This is impossible to verify; I included any picture that clearly shows the face of a person who I cannot identify as a celebrity.

Alonzo, Mei, and Milam Aiken. 2004. "Flaming in electronic communication." *Decision Support Systems* 36 (3): 205–213.

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.".

Coleman, LH, CE Paternite, and RC Sherman. 1999. "A reexamination of deindividuation in synchronous computer-mediated communication." *Computers in Human Behavior* 15 (1): 51–65.

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.

Dinakar, Karthik, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of Textual Cyberbullying. In *The Social Mobile Web*.

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.

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* .

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, deindividuation 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.".

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.

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.

Postmes, Tom, and Russell Spears. 1998. "Deindividuation and antinormative behavior: A meta-analysis." *Psychological Bulletin* 123 (3): 238.

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.

Roeckelein, Jon E. 1998. *Dictionary of theories, laws, and concepts in psychology.* Greenwood Publishing Group.

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.

Vandebosch, Heidi, and Katrien Van Cleemput. 2009. "Cyberbullying among youngsters: Profiles of bullies and victims." *New media & society* 11 (8): 1349–1371.

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

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*.

Ybarra, Michele L, Danah Boyd, Josephine D Korchmaros, and Jay Koby Oppenheim. 2012. "Defining and measuring cyberbullying within the larger context of bullying victimization." *Journal of Adolescent Health* 51 (1): 53–58.

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.