Don’t @ Me: Experimentally Reducing Partisan Incivility on Twitter

Kevin Munger*

February 16, 2017

Abstract

I conduct an experiment which examines the differential impact of several forms of moral suasion on partisans engaged in uncivil arguments. Partisans often respond in vitriolic and unproductive ways to tweets from politicians they disagree with, and this often engenders hateful responses from supporters of that politician. This phenomenon was especially common during the contentious 2016 US Presidential Election. Using Twitter accounts that I controlled, I sanctioned people engaged in this kind of behavior in October 2016. I found that two different forms of moral suasion were effective in decreasing incivility among conservatives, but that neither was effective in changing the behavior of liberals. These effects persisted for up to a month after treatment. My results suggest that it is possible for even to discourage political incivility and promote norms of polite discourse online.

1 Introduction

There is a general impression that political discourse today is less civil than it was twenty years ago, and that changing norms related to the style of cable news is partially

*Department of Politics, New York University, 19 West 4th Street, 2nd floor, New York, NY, USA. email: km2713@nyu.edu.
responsible (Berry and Sobieraj, 2013; Mutz, 2015). Scholars and the public alike are concerned that this incivility degrades the quality of political discussion and can cause people to withdraw from engaging in politics.

Even more recently, social media has become a major platform for news distribution, elite communication and mass political discussion. Although there is less scholarly work examining the impact of online civility on the political process (but see Theocharis et al. (2015)), there is a huge amount of concern about online harassment. The harassment of women and minorities, for example, can have a chilling effect on their online participation (Henson, Reyns, and Fisher, 2013; Hinduja and Patchin, 2007; Mantilla, 2013). This phenomenon has sufficiently permeated the Western cultural consciousness that no less august an institution than the long-running television series South Park dedicated an entire season to exploring the de-mobilizing effects of online harassment.

South Park was one of many observers to tie this trend to the US presidential candidacy of Donald Trump. Trump frequently decried “political correctness,” and many of his supporters cited his willingness to say exactly what he thinks without regard for violating speech norms as a major part of his appeal. Some of his statements were widely regarded as offensive to women, racial and religious minorities, and the disabled. Publicizing these views may have emboldened others who share them to express similarly offensive content online and led to a concerted effort by members of the “alt right” to harass people with whom they disagreed (Gross, 2016; Posner, 2016).

A similar trend has been observed with people on the opposite end of the political spectrum, sometimes called the “intolerant left.” Indeed, a major topic of discussion immediately before Trump’s campaign began dominating the headlines was the campus culture created by students’ attempts to dictate the range of tolerable discourse in the name of promoting social justice (Lukianoff and Haidt, 2015). Although their aims, tactics and moral frameworks differ from those on the alt right, online social justice activists have also engaged in coordinated efforts to harass people with whom they disagree (Ronson, 2016).

In this paper, I attempt to categorize and measure these behaviors in the context of the 2016 US Presidential election. I test different methods for encouraging civility online and evaluate their effectiveness. Using a method developed in an earlier paper (Munger, 2016), I used Twitter accounts that I controlled to sanction users engaged in uncivil discussions on Twitter. I sample users by searching for tweets that mention either @realDonaldTrump or @HillaryClinton but which are directed at another, non-elite user. Using an algorithm developed to identify aggression in comments on a Wikipedia
editors’ discussion forum (Wulczyn, Thain, and Dixon, 2016), I selected the tweets most likely to be uncivil. I then manually inspected the interaction to ensure that it was a true instance of a non-elite\textsuperscript{1} being uncivil to another non-elite of an opposing political persuasion. I then randomly assigned the subject to a treatment arm—subject to balance constraints—and used “bots” to send them a message.

By manipulating the partisan identity of my “bots”,\textsuperscript{2} I test the differential effects of sanctioning on Republicans and Democrats, as well as the degree of overlap between Democrats/Hillary supporters and Republicans/Trump supporters. By varying the language I tweet at subjects, I test hypotheses about the relative effectiveness of two kinds of moral suasion and include a “placebo check” by including a treatment arm with a message with no sanctioning.\textsuperscript{3}

I found evidence of significant changes in subjects’ behavior, but the effect heterogeneity took an unexpected form. There was no difference between the effectiveness of the two kinds of moral suasion, but there was a significant difference between liberal and conservative subjects: conservatives significantly reduced their rate of incivility in response to either treatment, but liberals did not change their behavior for either. This difference can be partially explained by much larger ideological heterogeneity among the liberal subjects. Additionally, although subject anonymity significantly moderated treatment effects, this moderation was in an unexpected direction: more anonymous subjects were less likely to respond to the treatment.

Overall, these findings demonstrate that various different forms of moral suasion can be effective in promoting a more civil political discourse on Twitter.

2 The Promise and Perils of Social Media

Perceptions of the impact of social media (and the internet more generally) on democratic politics have changed dramatically in the brief period of social media’s existence. Initial optimism suggested that citizens would be better able to communicate with both their governments and with each other, unconstrained by geography and the power im-

\textsuperscript{1}I define as an “elite” anyone who was “Verified” on Twitter—they have a blue check mark next to their name which means that Twitter has verified that they are who they say there, a status which Twitter only bestows on users they consider public figures—or anyone who identified themselves as a journalist or political operative in their profile.

\textsuperscript{2}These are not “bots” in the sense that they behave autonomously: I did all of the tweeting manually. I refer to them as bots throughout the paper for lack of a better term.

\textsuperscript{3}The research design, dependent variable measurement, and main hypothesis were pre-registered at EGAP.org prior to any research activities.
balances of the physical world (Papacharissi, 2002). Although conversations could get heated and impolite, the overall effect was to revitalize the public sphere of debate (Papacharissi, 2004). The campaign manager for Howard Dean, one of the first politicians in the US to fully embrace the power of the Internet for politics, said that “the Internet is the most democratizing innovation we’ve ever seen, more so even than the printing press” (Trippi (2004), quoted in Hindman (2008)).

Indeed, a wide variety of politicians began using social media to communicate with their constituents (Gulati and Williams, 2010). Individual politicians are better able to reach voters directly, rather than through the mediating institution of party control (Karlsen and Skogerbø, 2013). Although the process does not always work perfectly, there is evidence that politicians respond to the citizens who engage with them on social media, discussing topics that citizens bring to their attention (Barberá et al., 2014). Additionally, citizens do seem to learn about party platforms directly from communication by politicians on Twitter (Munger et al., 2016).

On the non-elite side, the use of the internet to discuss non-political topics has enabled some cross-cutting ideological mass discussion (Wojcieszak and Mutz, 2009). This phenomenon first began with blogs. By 2006, 8 million US citizens claimed to share their thoughts through online blogs, and fully 57 million US citizens claimed to read them (Hindman (2008), p104). Hindman describes the prevailing mood at that time, when media commentators were lauding the development of blogs as a brave new world for deliberative democracy: “The central claim about blogs is that they amplify the political voice of ordinary citizens.” However, as he argues persuasively in The Myth of Digital Democracy, the infrastructure of the internet tends to lead to an even more skewed distribution of readership than does traditional media: “It may be easy to speak in cyberspace, but it remains difficult to be heard. (p142)”

When the competition to be heard is intense, competitors often resort to using outrageousness to garner attention. For example, when cable enabled new entrants to the television marketplace, these upstart media organizations were willing to blend news and entertainment in a way that traditional network broadcasters had resisted. In the words of Bill O’Reilly, host of the famously confrontational television program The O’Reilly Factor: “The best [cable news] host is the guy or gal who can get the most listeners extremely annoyed over and over and over again” (O’Reilly (2003), cited in Mutz (2015)). Norms of journalistic integrity established in the early 20th century rapidly eroded, resulting in less civil media and citizens who trusted and liked that media less (Berry and Sobieraj, 2013; Ladd, 2011).
A similar trend took place in citizen online engagement, but more rapidly and to a greater extreme. Early forums tended to be anonymous, and early internet users flocked to sites like 4chan and somethingawful to discuss whatever was on their mind. However, a subset of these people found that this anonymity empowered them to say uncivil and outrageous things, and that they could easily upset other users. This behavior soon spread over the internet, as “trolls” mocked memorial pages on Facebook and posted vivid images of gore and hardcore pornography so that other users might suffer serious emotional turmoil (Phillips, 2015).

This kind of behavior is only possible through Computer Mediated Communication (CMC). In the physical world, biological feedback mechanisms make it emotionally difficult to look a stranger in the eye and say something uncivil (Frijda, 1988), but these mechanisms are lacking in CMC, as are physical proximity and identifiability. CMC makes it difficult to enforce social norms, and while this does tend to encourage more communication and creativity, it also allows even a small number of ill-intentioned actors to impose significant emotional costs on other users (Bordia, 1997; Kiesler, Siegel, and McGuire, 1984; Walther, 1996).

The competition for attention and the difficulty of punishment in anonymous contexts meant a race-to-the-bottom in terms of online speech norms, and the Internet is widely regarded as rife with offensive and even harassing speech designed to mock sincere expression—trolling culture is dominant online (Buckels, Trapnell, and Paulhus, 2014; Milner, 2013). The extent to which trolling culture obtains, though, depends on the specific technical affordances of different online platforms. The most important feature, in this respect, is the extent to which platforms allow their users to be anonymous. Studies have consistently found that the more anonymous platforms experience more harassment (Hosseinmardi et al., 2014; Omernick and Sood, 2013).

Facebook, for example, has invested heavily in linking their users’ accounts with their real identities. Twitter, on the other hand, allows all manner of parody, comedy and anonymous accounts. Twitter has consistently defined itself as in favor of free speech, and while this has made it the preferred platform for revolutionaries in both Western countries and authoritarian regimes around the world (Barberá et al., 2015; Earl et al., 2013), it has also become notorious for failing to curtail harassment. In the candid words of Twitter’s CEO Dick Costelo in an internal memo in 2015, “We suck at dealing with abuse and trolls on the platform and we’ve sucked at it for years.”
3 Social Media and Affect Polarization

The development of social media as both a platform for political communication and a locus for incivility took place at the same time as a sharp growth in animosity between Democratic and Republican partisans. Scholars have described this trend as “affect polarization”–partisans dislike each other (Iyengar, Sood, and Lelkes, 2012) and tend to trust co-partisans and distrust out-partisans more (Iyengar and Westwood, 2015). This phenomenon has even extended to the marriage market, as preferences for a partner with similar partisan characteristics is stronger than ever (Huber and Malhotra, 2013).

Although the uptick in partisan polarization began well before the mass adoption of social media, there exists a plausible connection between the two. Some scholars claim that social media use exposes people to a wider range of views and thus decreases issue polarization (Barberá, 2014), but others argue that social media inflames partisan emotions and increases affect polarization (Settle, Forthcoming). The large-scale, contemporaneous development of social media and affect polarization makes causal claims difficult to establish; an exception is Lelkes, Sood, and Iyengar (2015), who use the quasi-random rollout of broadband internet as an instrument for the use of social media and finds that it significantly increased affect polarization.

In some ways, incivility is entailed by increasing affect polarization. I follow Mutz (2015), who draws a connection between civility and following the norms of politeness in a given society: “Following the rules of civility/politeness is...a means of demonstrating mutual respect. (p7)” If mutual respect between partisans is decreasing, it should be no surprise that civility in their conversations is decreasing as well.

Regardless of causality, it is clear that uncivil political arguments take place on social media. Sometimes the incivility is directed at politicians themselves, and while we might expect that having a thick skin is necessary to survive in that business, Theocharis et al. (2015) show that this can decrease politician engagement with their constituents on Twitter. Perhaps more importantly for the mass public, this behavior means that citizens who wish to engage with politicians or each other in response to a politicians’ tweet are necessarily exposed to uncivil messages.

4 Experimentally Reducing Political Incivility

Although Twitter has made efforts to reduce the incidence of incivility and harassment, it remains a large problem. Building on previous work to experimentally reduce racist
harassment on Twitter (Munger, 2016), I conducted an experiment to sanction users who were sending uncivil messages to out-partisans and measured the change in their behavior.

The first step in performing this experiment was finding conversations that were uncivil, between out-partisans, and about politics. As in my previous experiment, where I searched for racist harassment by scraping tweets containing the slur “n****r,” I first attempted to use a keyword search. I could not figure out a term that would reliably find the interactions I was looking for.

Instead, I used streamR to scrape the streaming Twitter API for tweets mentioning either “@realDonaldTrump” or “@HillaryClinton”–the Twitter accounts of the two major party candidates in the 2016 US Presidential election. I then dropped any tweets that were not directed at another user who was not either Trump or Clinton. Sending an uncivil message to Twitter accounts managed by teams of campaign workers is not exactly morally laudable—it is perhaps akin to muttering obscenities at a campaign ad played on an airport television—but it is less important from a deliberative point of view.

This way, I found a sample of tweets from non-elites that were concerned with the “issues” most likely to inspire political incivility in October 2016: Trump and Clinton. In order to filter through the hundreds of thousands of tweets every hour that fit these criteria, I used a machine learning classifier developed by Wulczyn, Thain, and Dixon (2016) to detect aggression. Wulczyn and Thain trained and evaluated a neural network on millions of comments on Wikipedia “talk pages” (the behind-the-scenes part of Wikipedia where editors discuss potential changes) in a format that is reasonably similar in structure and length to Tweets.

I used the model to assign an “aggression score” to each tweet I had scraped, then manually evaluated the top 10% most aggressive tweets per batch. From these prospective subjects, I selected the ones who were directing incivil language at a member of the opposite political persuasion. Many of the potential subjects I found this way were tweeting at elites—either people verified on Twitter, journalists or campaign operatives—and I excluded them. I also found many people agreeing (though often in uncivil ways) with an in-partisan about how terrible the out-party is, and excluded them as well. When performing a manual inspection of the potential subject’s profile, I excluded users who appeared to be minors or who were not tweeting in English. I also checked

\[4\] This process was time-consuming, and there were a finite number of tweets satisfying my criteria being tweeted at a given time, so I iterated this scrape-validate-treat procedure several times.
to ensure that the subjects’ profile was at least two months old; Twitter does ban some user accounts for harassment or other violations of their Terms of Service, so a very new account is likely to have been started by someone who had previously been banned. A new user is also likely to have too short a tweeting history for me to establish a reasonable baseline for their past behavior.

For a visual overview of this selection process, see Figure 2. In this way, I found uncivil tweets from a non-elite to another non-elites with whom they disagreed politically. For an example, see Figure 1. @realDonaldTrump tweeted something, then Parker tweeted “you already lost” at Trump. Ty then responded to Parker (but because of how Twitter works, Ty’s tweet also “mentions” @realDonaldTrump) with an incivil comment. Ty is the subject I included in the experiment, and because he was being incivil to someone criticizing Trump, I coded Ty as a Trump supporter.

Based on findings in my previous experiment, and on the theoretical expectation that anonymity is an essential part of what enables incivility online, I also recorded each subject’s Anonymity Score during the subject discovery process. The Anonymity Score ranged from 0 (least anonymous, full name and picture) to 2 (most anonymous, no identifying information). Ty, from Figure 1, was coded as a 1–he chose to display what could plausibly be his full name. He also provided some personal information in his “bio” field, to the left of where he claims to be an “All around nice guy!”

My aim was to convince subjects that they were being sanctioned by a real person,

\[5\] I censor the usernames of the subjects to preserve their anonymity. In principle, the exact text of a tweet should be enough to find a user, but the phrases used in this exchange are quite common.
Figure 2: Sample Selection Process

StreamR finds a tweet with 
“@realDonaldTrump” or 
“@HillaryClinton”

Does the potential subject appear to be an adult speaking English, with a Twitter account at least 2 months old?

Is the incivility directed at someone besides a journalist or other political actor?

Is the incivility directed at someone who expressed a different political viewpoint?

Assign to a treatment condition subject to balance constraints

This flowchart depicts the decision process by which potential subjects were discovered, vetted and ultimately included or excluded.
so I made my bots look as real as possible. After I tweeted at a subject, they received a “notification” from Twitter. Non-elites are unlikely to get more than a few notifications per day, so they almost certainly saw the message I sent them. It is uncommon to be tweeted at by a stranger, but not extremely so, and especially not among a subject pool who are tweeting uncivil things at out-partisans. As a result, they were likely to click on my bots’ profile; if they did, they would see something very like Figure 3.

Neil, in panel (a), was a bot who appeared to be pro-Clinton. I created four bots; the other three were pro-Democrats, pro-Trump, and pro-Republicans (see Todd, in panel (b)). To manipulate these identities, I changed the large banner in the middle of the profile, the small logo in the bottom right of the bots’ profile pictures, and the “bio” field below their username (eg “Hillary 2016!”; “Republicans 2016!”). The four bots were otherwise identical. All of the bots appeared to be white men, keeping race/gender aspect of the treatment constant. I used identical cartoon avatars to avoid anything about the users’ appearance priming the subjects; it is not uncommon for Twitter users to have cartoon avatars, so this was unlikely to raise suspicions.

I took other steps in order to maximize verisimilitude. Most importantly, I ensured that all of the bots had a reasonably high number of followers. In Munger (2016), I varied the number of followers that sanctioning bots had, and found that bots with few followers had very little effect; several subjects even mocked the bots for having a low follower count. In the current experiment, I used the same “brand promotion” website to purchase 500 followers for each of my four bots, although each bot actually got 900 followers.\(^6\) The number did not vary significantly among the four.

I created each bot in January 2015, giving the impression that they were long-time users. When creating the accounts, I followed Twitter’s recommendation to follow 40 pre-selected accounts, mostly celebrities and news services. To further increase the perception that the bot was a real person, I tweeted dozens of innocuous observations (eg “I’m thinking of pasta for lunch.....YUM”) and retweeted random (non-political) stories from the accounts the bots followed.\(^7\)

The overall experimental setup is outlined in Table 1. There were two subject pools: people who were uncivil to people critical of Trump (“Conservatives”) and people who were uncivil to people critical of Clinton (“Liberals”). Within each of these pools,\

\(^6\)Interestingly, the price for 500 followers was $1 in Summer 2015, but the same website was charging $10 for the same service in Summer 2016. Other follower-selling sites had similarly increased their prices.

\(^7\)Bizarrely, the followers I bought sometimes “liked” and even occasionally retweeted these observations, suggesting that at least some of them are real people.
Figure 3: (a) Example Bot–Clinton Condition

(b) Example Bot–Republican Condition
Table 1: Experimental Design

<table>
<thead>
<tr>
<th></th>
<th>Conservative Subjects</th>
<th>Liberal Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trump Republicans</td>
<td></td>
</tr>
<tr>
<td>Feelings</td>
<td>$T_f$</td>
<td>$C_f$</td>
</tr>
<tr>
<td>Rules</td>
<td>$T_r$</td>
<td>$C_r$</td>
</tr>
<tr>
<td>Public</td>
<td>$T_p$</td>
<td>$C_p$</td>
</tr>
<tr>
<td></td>
<td>$R_f$</td>
<td>$D_f$</td>
</tr>
<tr>
<td></td>
<td>$R_r$</td>
<td>$D_r$</td>
</tr>
<tr>
<td></td>
<td>$R_p$</td>
<td>$D_p$</td>
</tr>
</tbody>
</table>

each subject was randomly assigned one of three messages (“Feelings”, (“Rules”, or (“Public”) sent by one of two bots (pro-candidate or pro-party). There were initially 118 subjects in the “Conservatives” pool, 104 subjects in the “Liberals” pool, and another 108 in the control group, to whom I sent no tweets.\(^8\)

The primary outcome of interest was how subjects responded to being sanctioned, both in terms of their direct response to the sanctioning tweet and in how they changed their behavior after having been sanctioned. I only used bots that appeared to be on the same “side” as subjects to send the sanctioning message; I was concerned that cross-ideological sanctioning might cause subjects to react angrily and send even more uncivil messages. I had no theoretical expectation as to whether right-leaning or left-leaning subjects would respond more to being sanctioned.

The primary variation in the treatments is in the language of the message sent to the subjects. The aim is to convince subjects that their behavior is wrong—or at a minimum, to convince them to change their behavior. One approach, the one I used in a previous experiment with bots in Twitter, is social norm promotion: to cause subjects to update their beliefs about correct normative behavior. I believe that this approach is less appropriate in the current context: as I argue above, incivility is well established as normatively acceptable in arguments on Twitter, and it is unlikely that a single intervention could cause subjects to update their views on this norm.

Instead, I attempted to use moral persuasion. I based my approach on the moral intuitionist model proposed by Haidt (2001), which argues that moral emotion is antecedent to moral reasoning. People make moral judgments based on deep-seated intuitions and then justify those judgments with ad hoc reasoning. As a result, moral appeals should be targeted to these fundamental intuitions, rather than to the putatively logical justifications for specific judgments.

Extending the theory, (Haidt, 2012) argues that a necessary component for moral

\(^8\)In the analysis below, I include 310 subjects out of this original pool of 330. I discuss the attrition process in Appendix A.
suasion is convincing your interlocutor that you are sympathetic and understanding. If
the two of you share the same fundamental moral intuitions, you can reasonably discuss
specific implications of those foundations, but if not, attempts to change their mind are
likely to be interpreted as attacks on their worldview and to be met with resistance. To
this end, all of my messages begin by identifying my bot and the subject as members
of the same party (Democrat/Republican).

Haidt also finds that the morality of liberals and conservatives rests on different
foundations. He finds six dimensions of morality that seem to operate in cultures
around the world: Care, Fairness, Liberty, Loyalty, Authority, and Sanctity. For an
action to fall in the realm of morality, it must either violate or uphold the principles
of these moral foundations. He argues that people in non-Western societies are similar
to conservatives in the West in that both groups appear to place significant weight on
all six of these moral foundations. Westerners on the left of the political spectrum,
however, appear to put far more emphasis on just two: Care and Fairness.

As a result, liberals and conservatives speak past each other on some moral issues.
For example, liberals sometimes have difficulty understanding why conservatives are
so upset about flag burning. Burning a flag does nothing to cause harm (the primary
question underlying the Care foundation), nor is it unfair, so liberals tend not to see
it in moral terms. Conservatives, though, feel that it is disloyal and disrespectful to
authority, and that flag burning is thus immoral.

To effectively engage in moral suasion, then, you must appeal to the correct moral
foundation of your interlocutor. To that end, I designed two different treatments. The
first was designed to appeal to the Care foundation, and thus to have some effect on
conservatives but a much larger effect on liberals:

@[subject] You shouldn’t use language like that. [Republicans/Democrats]
need to remember that our opponents are real people, with real feelings.

The other treatment appealed to the Authority foundation. My expectation was
that it should have an effect on conservatives but not on liberals:

@[subject] You shouldn’t use language like that. [Republicans/Democrats]
need to behave according to the proper rules of political civility.

In addition to these moral foundations treatments, I included a “placebo” treatment.
The goal was to separate out the effect of being tweeted at by a stranger from the specific
moral suasion of the main treatment tweets. My intention was to use a message that would serve to remind subjects that their uncivil tweets were public, and my hypothesis was that this treatment would decrease the subjects’ use of incivility, but that the effect would be smaller than the moral treatments. To that end, I designed a message that emphasized the subject’s visibility:

@[subject] Remember that everything you post here is public. Everyone can see that you tweeted this.

**Hypothesis 1** The reduction in incivility caused by the Care condition will be larger for liberals than for conservatives. There should be a reduction in incivility caused by the Authority condition for conservatives, but not for liberals. There should be a reduction in incivility caused by the Public condition, but it should be smaller than the other effects.

Some subjects are more heavily invested in their online identities than are others. Twitter allows individuals to decide how much personal information to divulge, so while some users are completely anonymous, others include their full name, picture, and biography. There are likely to be large differences in how these different types of users engage with Twitter. In my previous bot experiment, I found that more anonymous users were more likely to change their behavior in response to being sanctioned, and I expected the same to be the case here.\(^9\)

**Hypothesis 2** The reduction in incivility caused by the treatments will negatively co-vary with the subject’s Anonymity Score.

5 Results

The behavior targeted in this experiment is partisan incivility targeted at other Twitter user. To capture this behavior, I scraped each subject’s Twitter history before and after the treatment and restricted the sample to the tweets that were “@-replies”: tweets directed at another user. After removing the 18 users for whom I could not collect enough pre- or post-treatment tweets (see Appendix A for a full discussion), I used the model trained by Wulczyn, Thain, and Dixon (2016) to assign an “aggression score” (between 0 and 1) to each of these 367 thousand tweets. This measure was

\(^9\)Note that this hypothesis was not recorded in the Pre-Analysis Plan, but follows directly from the findings in Munger (2016).
skewed toward the lower end of the distribution, so I selected all tweets above the 75th percentile aggression score and coded them as uncivil.10

To control for each subjects’ pre-treatment behavior, I calculated their rate of uncivil tweeting in the three months before the experiment. This measure was included as a covariate in all of the following analysis. I then calculated this same measure for different post-treatment time periods, to test for effect persistence.

Because these are overdispersed count data, I used a negative binomial regression to calculate effects. The experimental results on the full sample are displayed in Figure 4.11 The Public and Rules treatments produced a statistically significant reduction in incivil tweets directed at another user. As expected, the Public treatment condition had an effect in the expected direction—although this effect is significant only in the middle panel—but it was smaller than the effect of the two moral treatments.

As predicted, the Anonymity Scores of the subjects significantly moderated the treatment effects. However, the effect was in the opposite direction: the effects were larger on the subjects with lower Anonymity Scores (who provided more information on their profiles). This was true across all three treatment conditions, although the effect was only statistically significant for the Rules condition and a short time period for the Feelings condition.

The negative binomial specification is estimated using the following model:

\[
\ln(Agg_{\text{post}}) = x_{\text{int}} + \beta_1 Agg_{\text{pre}} + \beta_2 T_{\text{feel}} + \beta_3 T_{\text{rules}} + \beta_4 T_{\text{public}} + \beta_5 \text{Anon} + \beta_6 (T_{\text{feel}} \times \text{Anon}) \\
\quad + \beta_7 (T_{\text{rules}} \times \text{Anon}) + \beta_8 (T_{\text{public}} \times \text{Anon})
\]

To interpret the relevant treatment effects implied by the coefficients estimated by this model, the exponent of the estimated \(\hat{\beta}_k\) for each of the treatment conditions needs to be added to the corresponding \(\hat{\beta}\) for the interaction term, evaluated at each level of Anonymity Score (Hilbe, 2008). For example, the effect of the Feelings treatment on subjects with Anonymity Score 1 (the middle category) is:

\[
IRR_{\text{feel} \times \text{Anon}1} = e^{\hat{\beta}_2 + \hat{\beta}_6 \times 1}
\]

10 Results are largely unchanged if I select the 75th or 85th percentile.
11 Results using OLS and a logged dependent variable are presented in Appendix C. The results are all substantively the same, although the time period in which effects remain statistically significant is shorter.
**Figure 4:** Each panel represents the results of a separate negative binomial regression in which the outcome variable is the absolute number of times a subject directed an uncivil tweet at another user in the specified time period. Each regression also controls for the log of the subject’s absolute rate of aggressive tweeting in the three months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.
These Incidence Ratios for the 7-day post-treatment period are plotted in Figure 5. $IRR_{feel \times Anon_1} = 0.43$ can be seen in the blue line in the middle of the plot. This Incidence Ratio implies that the average subject with Anonymity Score 1 who received the Feelings treatment tweeted 43% as many aggressive tweets as the average subject with Anonymity Score 1 in the control condition.\footnote{Note that this approach assumes that treatment effects are constant, and holds the pre-treatment level of aggressive treats constant at its mean level.} The confidence intervals in Figure 5 are calculated from the estimated variance of this estimator:

$$V_{feel \times Anon_1} = V(\hat{\beta}_2) + Anon^2 V(\hat{\beta}_6) + 2Anon \times Cov(\hat{\beta}_2, \hat{\beta}_6)$$

The Feelings and Rules treatments were both effective on subjects who shared at least some biographical information on their account, and the Feelings treatment caused a significant effect even on the fully anonymous subjects. The Public treatment, however, only had a significant effect on the least anonymous subjects. Note that these are ratios: going from .5 to 1 represents the same effect size (a 100% increase) as going from 1 to 2, so the upper half of the confidence intervals appear longer than the lower half.

To test Hypothesis 1, Figure 6 plots the same analysis on the populations divided into Anti-Trumpists (Liberals) and Anti-Hillaryites (Conservatives). I find very little support for my hypothesis that the Rules treatment would only be effective on Conservatives; in all six plots, the effect of the Rules and Feelings treatments are roughly similar. I also fail to find the expected greater effect for the Feelings treatment relative to the Rules treatment among Liberals; if anything, the effect of the rules treatment appears to be larger, the opposite of my expectations.

However, there was a large disparity in the overall effect of the experiment on Liberals and Conservatives. Although the effects of the Rules and Feelings treatments were in the expected direction, there was not a statistically significant effect on Liberal subjects in any time period.\footnote{The Rules treatment is significant at $p = .1$ in the far right panel, but not in the previous two panels. There is no plausible mechanism by which the treatment would have an effect only after two weeks, so this is likely to be spurious.} On the other hand, all three treatments had a significant effect on Conservatives; as the far right panel shows, this effect persisted for a month after treatment.

One possible explanation for the lack of an effect on Liberals is that this group was more heterogeneous. I implemented the method developed by Barberá (2015) to
Figure 5: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the 7-day post-treatment period. For example, the Incidence Ratio associated with the Feelings treatment on subjects with Anonymity Score 1 in the middle of the plot means that these subjects sent 43% as many directed uncivil tweets as the subjects with Anonymity Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.
Liberals \((N=147)\)

Conservatives \((N=163)\)

Figure 6: Each panel represents the results of a separate negative binomial regression in which the outcome variable is the absolute number of times a subject directed an uncivil tweet at another user in the specified time period. The top three plots are calculated only on the Liberal sample, and the bottom three plots only the Conservative sample. Each regression also controls for the log of the subject’s absolute rate of aggressive tweeting in the three months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.
estimate subjects’ ideological ideal points. This was possible for 314 of the 330 subjects who followed enough American political elites. As Figure 7 demonstrates, there was significant heterogeneity in the ideal points of subjects I coded as Liberals, but not for Conservatives.

All but two of the subjects coded as Anti-Hillary (Liberals) had estimated ideology scores above 1, and only one was coded as left of center. However, a full third of the subjects coded as Anti-Trump (Conservatives) had estimated ideology scores right of center, although only a few are far to the right (have an ideology score above 1). Looking at Figure 7, there appears to be two distinct clusters of Anti-Trump subjects; it may be that there was a significant contingent of moderate Anti-Trump Conservatives that I classified as Liberals. Because the Feelings and Public treatment messages were explicitly designed to appeal to subjects’ partisan group identities (and identified the anti-Trump subjects as “Democrats”), this ideological heterogeneity could pose a problem for estimating average treatment effects.

If I restrict the analysis of Liberals in Figure 6 to only those with estimated ideology scores to the left of center, I find some support for this _ex post_ explanation. While the
point estimates of the effect sizes are roughly double the point estimates in the full sample, the results remain insignificant because of the reduced sample size. Additionally, these differences are only noticeable for the Feelings and Rules treatments; for the Rules treatment, which did not explicitly refer to the subjects as Democrats, there is a much more modest effect. See Appendix B for the presentation of these results.

6 Conclusion

The 2016 US Presidential Election took place in the context of a deeply polarized electorate. Many partisans refrain from engaging in political discussion in their day-to-day lives for fear of alienating members of their communities: Berry and Sobieraj (2013) performed dozens of in-depth interviews with partisans who explained that they often self-censored to “avoid offending others or engaging in awkward social exchanges.” However, the authors noticed an asymmetry between liberal and conservatives—“conservative respondents alone...[fear] being judged negatively as people because of their view” (emphasis in original).

This offers an explanation for the main (unexpected) result from the experiment discussed in this paper: social sanctioning from Twitter bots was much more effective at causing conservatives to decrease their rate of uncivil tweeting. The difference in the effects on conservatives and liberals becomes smaller when I remove subjects classified as liberal who may actually have been right of center, but a large gap remains. There was also little difference between sanctioning language designed to appeal to subjects’ moral sense of Care or Authority. I expected the former to be more effective on liberals and the latter on conservatives, but there was little effect heterogeneity.

Part of the lack of a response from liberals to the Feelings treatment may be explained by the tweets then sent to my bots in response to being sanctioned. In several cases, liberals told my bots something like “these other people are Trump supporters, so I don’t care about their feelings”; no conservatives expressed a similar sentiment. The Trump campaign elicited extremely strong reactions from some liberals, so it is possible that this liberal resistance to moral suasion based on Care was idiosyncratic to the 2016 election.

Another insight from Berry and Sobieraj (2013)’s partisan interviews is that this restraint from talking contentious politics might be context-specific: one subject “[wasn’t] rattled by social conflict, as she is comfortable being politically contrarian under the
cloak of anonymity.” The subjects in my study may have felt similarly: contrary to my expectations, the treatment effects were largest on the subjects who were the least anonymous.

This was particularly surprising because the subjects’ anonymity played a significantly different role in a previous experiment using Twitter bots to sanction users engaged in racist harassment. The role of anonymity in moderating how people engage in online communication is a complicated one, but in the context of Twitter, a semi-anonymous platform in which each user can select her own level of anonymity, these moderating effects are likely to signal differences in the type of user rather than the impact of anonymity *per se*.

My post-hoc explanation for the inverted relationship between anonymity and treatment effectiveness in the two studies comes from the composition of the subject pool in each case. Among people using racist slurs, the ones who provided a full biography were fully committed to and unashamed of this behavior, and the treatment was more effective on the more impressionable anonymous users who were aware that their behavior was wrong. The behavior sanctioned in the current study, sending uncivil tweets at partisans from the other side, is almost certainly less objectionable than tweeting racist slurs. The fully anonymous users in this sample, then, may have been more likely to be committed “trolls” than normal (if passionately polarized) people. This explanation concords with the estimated magnitudes of the effects: the Rules treatment was more effective than the Feelings treatment among the least anonymous users, but the Rules treatment actually caused an (nonsignificant) increase in incivility among the most anonymous users.

This finding fits in with recent research on online trolling, and suggests a way to improve online discourse. Cheng et al. (2017) finds that there are a small number of dedicated online trolls, but that a much larger group of people will use uncivil language on forums where others have already been uncivil. These are precisely the people who may constitute the subject pool of this experiment: they saw others say something nasty to their preferred candidate, and responded in kind.

It may be difficult to prevent hardcore trolls from setting an uncivil tone, but my findings suggest that it may be possible to prevent incivility from becoming the norm by reminding normal people of our shared humanity and responsibility to the rules of civil discourse.
Appendix

A Attrition

Although I initially recorded 330 subjects as belonging to either a treatment or control condition, the final analysis includes only 310 subjects. The sample suffered from attrition from one of four sources.

In the case of four subjects, I mis-applied the treatment. When I used my bots to tweet at the subjects, I made a computer error and tweeted directly at them rather than in response to a specific uncivil tweet. I became aware of this possibility when one subject responded to my tweet in confusion; in re-checking the rest of the subjects, I found the other 3 mistakes.

I identified the rest of the potentially problematic subjects through patterns in their tweeting behavior. I manually re-inspected all of the profiles of subjects for whom I collected fewer than 50 tweets pre-treatment and 50 tweets post-treatment. The majority of the profiles I identified this way still merited inclusion; they were just people who did not tweet very often. However, I excluded others from the final sample. I did this manual re-inspection before calculating any of the results and without knowledge of which treatment condition to which the subjects belonged.

The most common problem was that I had 0 pre-treatment tweets for a subject despite having thousands of post-treatment tweets. This was caused by the timing of when I scraped their profiles and the Twitter API’s historical tweet limit: Twitter will only give you the 3,200 most recent tweets from a given account. I performed a full scrape of each account within a week of the treatment; this implies that these accounts were tweeting thousands of times a week. This is very difficult for a human to do, so I suspect that many of these accounts were bots; if they were not bots, they were extremely atypical Twitter users. However, this was the single largest source of attrition; just under 3% of the original accounts were excluded for this reason.

There were a total of 3 accounts in my sample that were suspended by Twitter during the course of my experiment. I do technically have enough tweets from these accounts to include them in the analysis, but doing so has the potential to bias my results upwards: the reduction in the number of uncivil tweets they sent was actually caused by Twitter preventing them from tweeting, rather than by the treatment.

Finally, there were two accounts that were just weird; they had not tweeted thou-
Table 2: Attrition Rates and Causes

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Liberals</th>
<th>Conservatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial assignment</td>
<td>108</td>
<td>104</td>
<td>118</td>
</tr>
<tr>
<td>Failed treatment application</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Tweeted too often/bots</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Suspended</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Weird</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Final</td>
<td>102</td>
<td>100</td>
<td>108</td>
</tr>
<tr>
<td>Attrition</td>
<td>6%</td>
<td>4%</td>
<td>8%</td>
</tr>
</tbody>
</table>

sands of times, but each still only recorded 3 pre-treatment tweets. In both cases, the accounts appeared to be behaving very oddly, and since I did not have a reasonable estimate of their pre-treatment behavior, I excluded them.

B Restricted Analysis of Accurately Classified Liberals

As shown in Figure 7, 61 out of the 147 subjects classified as “Liberals” because they sent anti-Trump tweets were actually estimated to have ideology scores to the right of center, using the method developed by Barberá (2015). The lack of any significant effects on the entire Liberals sample may have been caused by conflating a broadly heterogeneous group of subjects under this heading. Table 3 lists the point estimates of the main treatment effects.

The left column presents the point estimates plotted in Figure 6 for the full Liberals sample, while the right column presents the point estimates if the sample is restricted to those 86 subjects with ideologies estimated to be left of center.

C OLS Specification of Main Results

The dependent variable of interest in this analysis is the number of times a subject sent an uncivil tweet to another user. This is a “count variable”—it can only take non-negative integer values—and thus violates a fundamental assumption of OLS regression. To address this issue, generalized linear models with different assumptions are often used. Poisson regression, in which the dependent variable is assumed to have a Poisson
distribution, is a common technique, but this carries the further assumption that the variance and expected value of the dependent variable are equal. In cases in which the variance is significantly higher than the expected value–like it is here–the negative binomial model relaxes this assumption (Hilbe, 2008).

This means the negative binomial model used in the body of the paper contains assumptions about the shape of the distribution of the outcome variable as well, and there are some scholars who believe that the potential bias generated by violations of assumptions of parametric models like these pose a greater risk than that of straightforward OLS regression. To address this possibility, I re-ran the analysis in the body of the paper using OLS, using the log of the number of incivil tweets as at the dependent variable.

The results in Figure ?? are very similar to those in Figure 4. The point estimate for the Rules treatment is largest, followed by the Feelings treatment and then the Public treatment; the former two are statistically significant in the 1 day period. They are just shy of significance at $p < .1$ in the longer time periods, while the specification in Figure 4 suggest significant effects that persist.

The bottom row of Figure ?? shows the same analysis but with the 61 misclassified Liberals (discussed in Appendix B) removed. The point estimates of the effects are larger in magnitude in the bottom row, and both the Feelings and Rules treatments
have significant effects in the 2-7 day time period, even as the reduced sample size results in larger standard errors.

The overall inferences from the negative binomial regressions run in the body of the text are robust to using OLS. The models disagree about whether the effects of the Feelings and Rules treatments persist for the 15-28 day time period; my belief is that the negative binomial regression is the correct model, but researchers might reasonably disagree and assign less credibility to the persistence of the effects.
Figure 8: Each panel represents the results of a separate OLS regression in which the outcome variable is the log of the number of times a subject directed an uncivil tweet at another user in the specified time period. The top three plots are calculated only on the Liberal sample, and the bottom three plots only the Conservative sample. Each regression also controls for the log of the subject’s absolute rate of aggressive tweeting in the three months prior to the treatment. The vertical tick marks represent 90% confidence intervals and the full lines represent 95% confidence intervals.
References


Gulati, Jeff, and Christine B Williams. 2010. “Communicating with constituents in 140 characters or less: Twitter and the diffusion of technology innovation in the United States Congress.” Available at SSRN 1628247.


Huber, Gregory, and Neil Malhotra. 2013. Dimensions of political homophily: Isolating choice homophily along political characteristics. In American Political Science Association annual meeting, New Orleans, LA.


Phillips, Whitney. 2015. *This is why we can’t have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.


