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

Kevin Munger*

April 3, 2017

Abstract

I conduct an experiment which examines the impact of moral suasion on partisans engaged in incivil arguments. Partisans often respond in vitriolic ways to politicians they disagree with, and this can engender hateful responses from partisans from the other side. This phenomenon was especially common during the contentious 2016 US Presidential Election. Using Twitter accounts that I controlled, I sanctioned people engaged partisan incivility in October 2016. I found that two different forms of moral suasion were effective in decreasing incivility among Republicans, but that neither was as effective in changing the behavior of Democrats. These effects were significantly moderated by the anonymity of the subjects, especially among Republicans, many of whom may have been committed trolls. In some cases, the reduction in incivility persisted for up to a month after treatment. My results suggest an avenue for discouraging political incivility and promoting norms of less polarizing discourse online.

*Department of Politics, New York University, 19 West 4th Street, 2nd floor, New York, NY, USA. email: km2713@nyu.edu.

1 Introduction

In October of 2016, President Obama claimed (and Democratic Presidential nominee Hillary Clinton tweeted) that "civility is on the ballot." Concern over political civility was widespread during the 2016 US Presidential election, and many felt that the internet and social media (which Republican Presidential nominee Donald Trump employed enthusiastically) were to blame.

The trend towards incivility in contemporary political discourse can be traced back at least to the rise of cable news and its personalistic, outraged style (Berry and Sobieraj, 2013; Mutz, 2015). Indeed, concern about civil discourse may accompany any technological advance that lowers the cost of information production and distribution; the invention of the printing press led to elite concern about civil discourse during the time of the Reformation (Bejan, 2017).

Modern technological changes are taking place in the context of increased partisan animosity. Often called "affect polarization," this animosity reflects a growing distrust and lack of respect among Democrats and Republicans (Iyengar, Sood, and Lelkes, 2012). This phenomenon is directly related to civility, which Mutz (2015) says is "a means of demonstrating mutual respect" (p7). Incivility is more than impoliteness: it is indicative of a disregard for the act of deliberation. Internet technologies may or may not be driving affect polarization, but they do at a minimum allow for the lack of mutual respect to manifest itself in incivil online discourse.

Online communication lacks the biological feedback that makes it difficult to be incivil in a real-world setting, and it affords physical distance and (sometimes) anonymity, decreasing the effectiveness of social sanctioning (Frijda, 1988). These technological affordances, in a context replete with bad actors intent on sowing discord for fun (Phillips, 2015) or geopolitical advantage (Chen, 2015), have degraded norms of civil discourse online.

The implications of these changing norms are serious. Early enthusiasm about the capacity of the internet to democratize political discourse may have been premature (Hindman, 2008), but the affordances of today's ubiquitous, easy-to-use and social internet have caught up with the hype: in 2008, only 25% of US adults were on a social network, but during the 2016 US Presidential Campaign, that number was 68% (Greenwood, Perrin, and Duggan, 2016).

Ideally, the internet could enable broad, direct, deliberative democracy, with all of the desirable normative qualities this entails. Technological advances may continue to

expand the breadth and depth of online communication, but if incivil discourse remains the norm, deliberative democracy will remain out of reach.

Deliberative democracy entails more than mere communication. It only works to the extent that participants sincerely weigh the merits of the arguments being deliberated and that this consideration is not contingent on the identity of the person making the argument (Fishkin, 2011). In a context of high affect polarization and norms of partisan incivility, this rhetorical charity does not obtain. Survey evidence suggests that internet users do not feel that their interactions are deliberative–"64% say their online encounters with people on the opposite side of the political spectrum leave them feeling as if they have even less in common than they thought" (Duggan and Smith, 2016).

I conducted an experiment that evaluates different strategies for promoting civil political discourse during the 2016 US Presidential election. Using a method developed in an earlier paper (Munger, 2016), I used Twitter accounts that I controlled to sanction users engaged in incivil discussions. In contrast to lab experiments conducted on a convenience sample in a short time frame, this approach allowed me to measure the effectiveness of sanctioning on a sample of frequently incivil partisans in a realistic setting and in a continuous and unbounded time frame.

Users were sampled by searching for tweets that mentioned either @realDonaldTrump or @HillaryClinton but which were 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 incivil. I then manually inspected the interaction to ensure that it was a true instance of a non-elite[1] being incivil to another non-elite of the opposing partisan 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,"[2] I test the differential effects of sanctioning on Republicans and Democrats. By varying the language I tweeted at subjects, I test hypotheses about the relative effectiveness of two kinds of moral suasion

[1]I define as an "elite" anyone who was "Verified" on Twitter–they had 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.

[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.

and include a "placebo check" of a message with no moral sanctioning.[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 Democrat and Republican subjects: Republicans significantly reduced their rate of incivility in response to either treatment, but Democrats did not change their behavior for either. This difference can be partially explained by much larger ideological heterogeneity among the Democrat 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.

These findings demonstrate that various different forms of moral suasion can be effective in promoting a more civil political discourse on Twitter, above and beyond the effect of merely calling attention to the subjects' behavior. This moral suasion may only be effective on a subset of users; anonymous users (those more likely to be trolls) were unresponsive to moral suasion, and may even have been encouraged by being told that they were violating norms of political civility. Efforts to promote online civility should be sure to target the right people and use the most appropriate rhetorical strategy to maximize their efficacy.

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 imbalances 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

[3]The research design, dependent variable measurement, and main hypothesis were pre-registered at EGAP.org prior to any research activities.

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 incivil 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 incivil (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. Today, 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 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 Affect Polarization and Deliberation

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 find that it significantly increased affect polarization.

Regardless of causality, it is clear that incivil 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 incivil messages. The presence of incivility thus has a *compositional* effect on online political discourse: only people with a high tolerance for incivil discourse can hope to engage in public discussions. There also appears to be a *direct* effect of incivility on an individual's discursive style: Cheng et al. (2017) find that discussants who join an online forum and see that an incivil discussion is taking place are more likely to be incivil themselves. These two effects have allowed the norm of incivility in online discussions set by a small group of committed trolls to set the tone for a lot of online discussions.

Incivility comes far more naturally if you believe your interlocutor deserves it; in some ways, incivility is entailed by increasing affect polarization. I follow Mutz (2015): "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. The implications for deliberative democracy are serious; Fishkin's model claims that deliberative democracy works to the extent that participants sincerely weigh the merits of the arguments being deliberated and that this consideration is not contingent on the identity of the person making the argument (Fishkin, 2011). This does not at all describe the dominant mode of political discourse online: rather than leading to an exchange of information and arguments that can potentially lead to a consensus, a name-calling match between partisans online may

actually cause both parties to think less of their opponents and their arguments, driving the parties even further from consensus.

4 Experimentally Reducing Political Incivility

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

The first step in performing this experiment was finding conversations that were incivil, 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 incivil 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.

In 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.[4] From these prospec-

[4]This process was time-consuming, and there were a finite number of tweets satisfying my criteria

Figure 1: Finding Non-Elite Incivility

tive 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 incivil 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 to ensure that the subject's 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 establish a reasonable baseline for their past behavior.

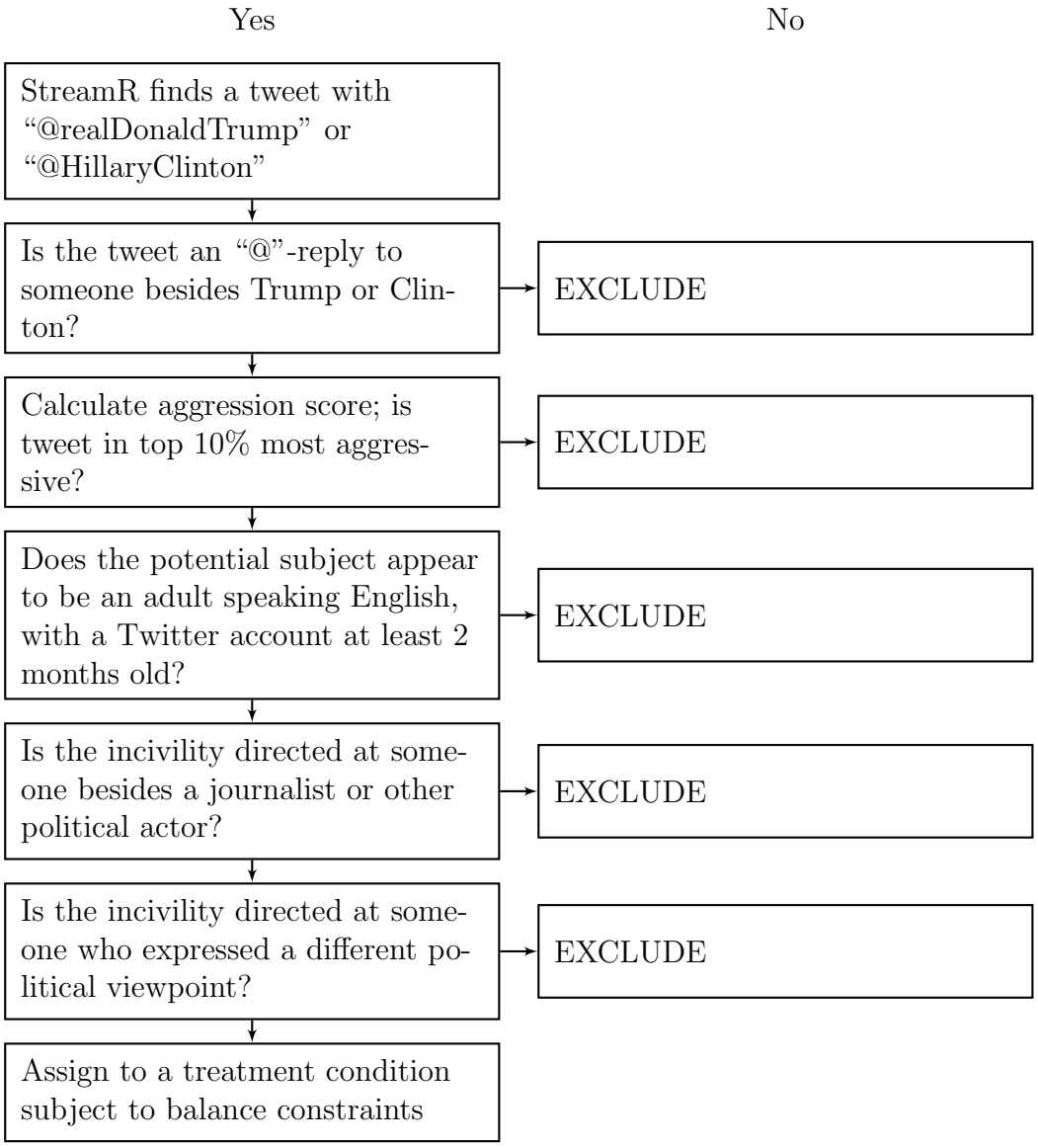
For a visual overview of this selection process, see Figure 2. In this way, I found incivil tweets from a non-elite to another non-elite with whom they disagreed politically. For an example, see Figure 1. @realDonaldTrump tweeted something, then Parker tweeted "you already lost" at Trump.[5] Ty then 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

being tweeted at a given time, so I iterated this scrape-validate-treat procedure several times.

[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



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

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!", which I censor for privacy reasons.

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

[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.

Figure 3: (a) Example Bot–Clinton Condition

(b) Example Bot–Republican Condition

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]

There were two subject pools: people who were incivil to people critical of Trump ("Republicans") and people who were incivil to people critical of Clinton ("Democrats"). Within each of these pools, 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 "Republicans" pool, 104 subjects in the "Democrats" 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 incivil 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 on Twitter, is *in-group social norm promotion*: to cause subjects to update their beliefs about correct normative behavior for someone sharing their social identity. I found that sanctioning from bots that shared a social identity with the subject were more effective in changing their behavior than bots with a different social identity. To build on this finding, I held in-group social identity (in this case, partisanship) constant in the current study.

By varying the language of the in-group sanctioning, I tested the possibility of moral suasion. 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

[7]Bizarrely, the followers I bought sometimes "liked" and even occasionally retweeted these observations, suggesting that at least some of them are real people.

[8]In the analysis below, I include 310 subjects out of this original pool of 330. I discuss the attrition process in Appendix A.

intuitions, rather than to the putatively logical justifications for specific judgments.

Extending the theory, (Haidt, 2012) argues that a necessary component for moral 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 Republicans but a much larger effect on Democrats:

```
@[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 Republicans but not on Democrats:

```
@[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 incivil 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 Democrats than for Republicans. There should be a reduction in incivility caused by the Authority condition for Republicans, but not for Democrats. There should be a reduction in incivility cased 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 covary 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

[9]Note that this hypothesis was not recorded in the Pre-Analysis Plan, but follows directly from the findings in Munger (2016).

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 skewed toward the lower end of the distribution, so I selected all tweets above the 75th percentile aggression score and coded them as incivil.[10]

To control for each subjects' pre-treatment behavior, I calculated their rate of incivil 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 negative binomial regression. The negative binomial specification is estimated using the following model:

$$ln(Agg_{post}) = x_{int} + \beta_1 Agg_{pre} + \beta_2 T_{feel} + \beta_3 T_{rules} + \beta_4 T_{public} + \beta_5 Anon + \beta_6 (T_{feel} \times Anon)$$

$$+\beta_7 (T_{rules} \times Anon) + \beta_8 (T_{public} \times Anon)$$

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:

$$IRR_{feel \times Anon_1} = e^{\hat{\beta}_2 + \hat{\beta}_6 \times 1}$$

The experimental results on the full sample in the first day after treatment are displayed in Figure 4; in all of the analysis that follows, the dependent variable is the number of incivil tweets the subject sent in the specified time period.[11] The Public and Rules treatments produced a statistically significant reduction in incivil tweets directed at another user among subjects with Anonymity Score 0 or 1. As expected, the Public treatment condition had an effect in the same direction, but it was smaller than

[10]Results are largely unchanged if I select the 70th or 80th percentile. Because the treatment could affect the distribution of aggression scores, I looked only at pre-treatment tweets when calculating these percentiles.

[11]Results using OLS and a logged dependent variable are presented in Appendix B. The results are all substantively the same, although the time period in which effects remain statistically significant is shorter.

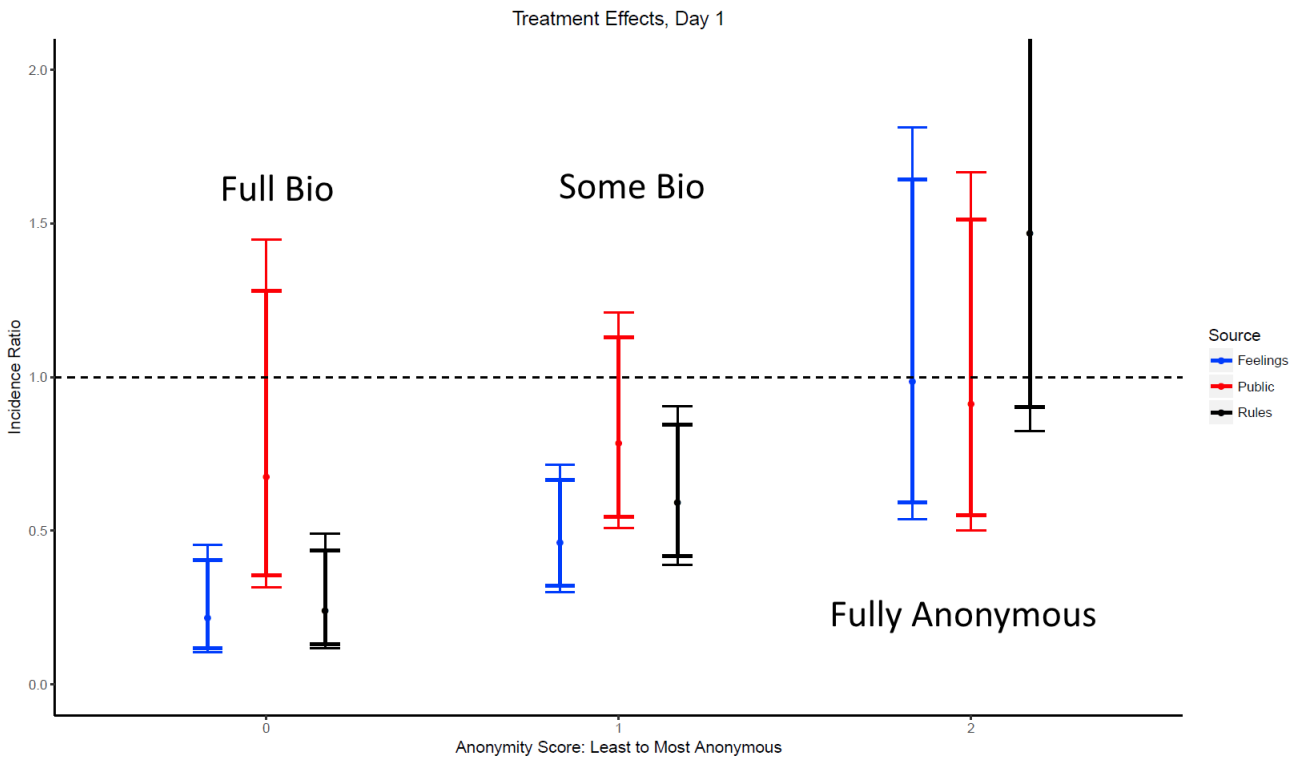


Figure 4: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first day after treatment. For example, the Incidence Ratio associated with the Feelings treatment on subjects with Anonymity Score 1 in the middle of the plot means that these subjects sent 45% as many directed incivil 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.

the effect of the two moral treatments. However, there was no statistically significant reduction in incivility among the fully anonymous subjects; in fact, the Rules treatment caused a (non-significant) increase in incivility among these subjects.[12]

$IRR_{feel \times Anon_1} = 0.45$ 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 45% as many directed incivil tweets as the average subject with Anonymity Score 1 in the control condition.[13] The confidence intervals in

[12]Because of the presence of these strongly heterogeneous effects, there are no significant treatment effects on the full sample without any interaction terms, so I do not show these results.

[13]Note that this approach assumes that treatment effects are constant, and holds the pre-treatment

Figure 4 are calculated from the estimated variance of this estimator:

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

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.

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 both moral suasion treatment conditions; there were no significant effects of the Public treatment.

To test for effect persistence, Figure 5 plots the results in the first week (excluding day 1; top panel) and third and fourth weeks (bottom panel) after treatment. The effects in the first week are substantively similar to those in Figure 4, except that the former shows a significant reduction for the Public treatment among the least anonymous subjects.

Most of the effects in the bottom panel of Figure 5, weeks 3 and 4 post-treatment, are no longer significant. The one exception is the Rules treatment on the least anonymous subjects. There is also a marginally significant *increase* associated with the Rules treatment among the most anonymous subjects; there is no mechanism to explain why the effect of this treatment would get stronger over time, so this result may be spurious.

To test Hypothesis 1, Figure 6 plots the same analysis on the populations divided into Anti-Trumpists (Democrats) and Anti-Hillaryites (Republicans) in the 2-7 day time period. There are fewer significant results in these models because the sample sizes have been halved, but the contrast is informative. Among Democrats, in the top panel, there were only significant treatment effects for the Feelings treatment, and only among partially or fully anonymous subjects.

Among Republicans, in the bottom panel, the results appear more similar to the results in Figure 4, although the Feelings treatment ceases to show an effect on the subjects with Anonymity score 1. The most striking result is that all three treatment conditions caused a (nearly significant) *increase* in directed incivility among the fully anonymous Republican subjects, with the Rules treatment again causing the largest increase.

---

level of aggressive treats constant at its mean level.

Change in Incivility, Treatment Persistence (*N=310*)

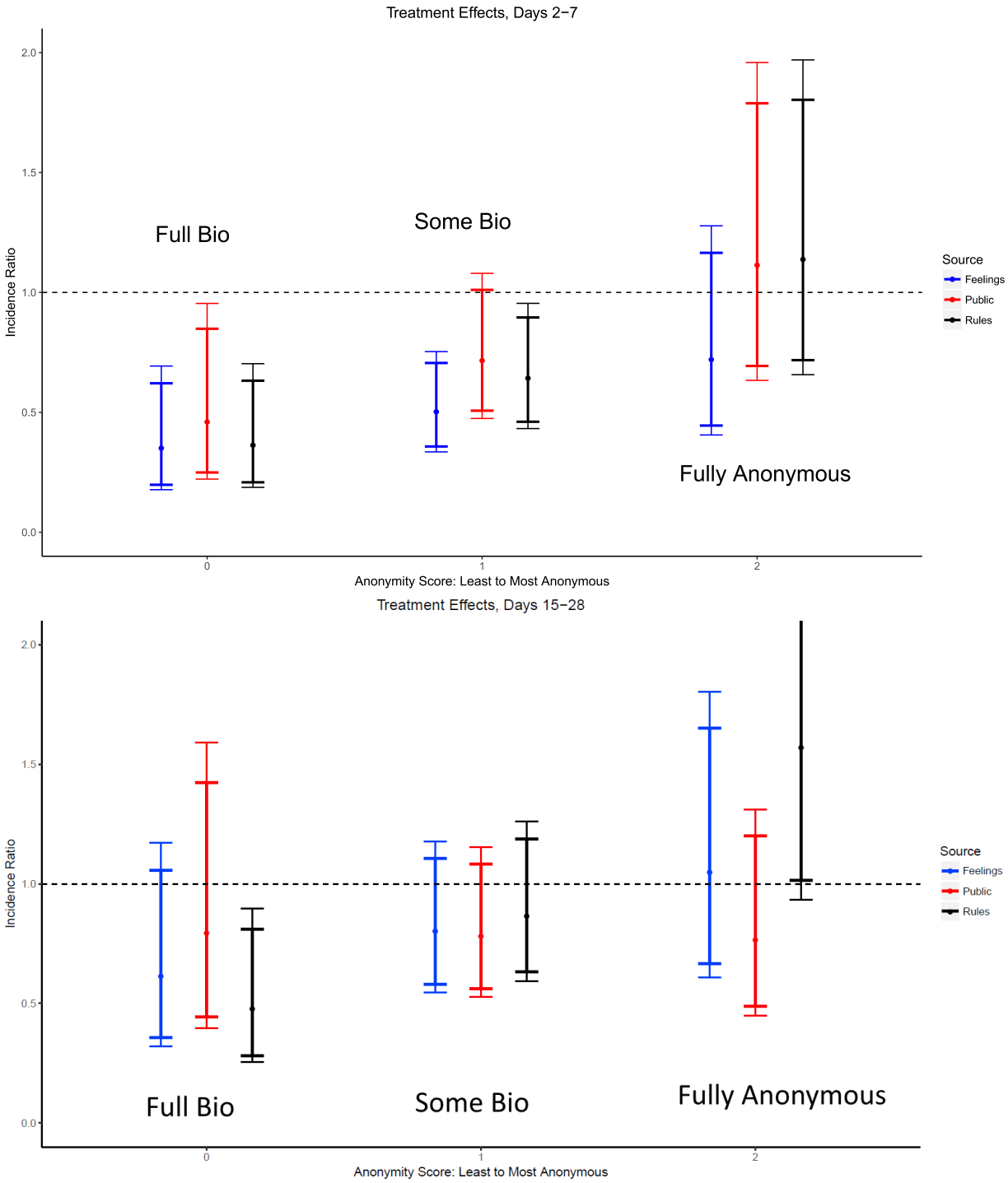
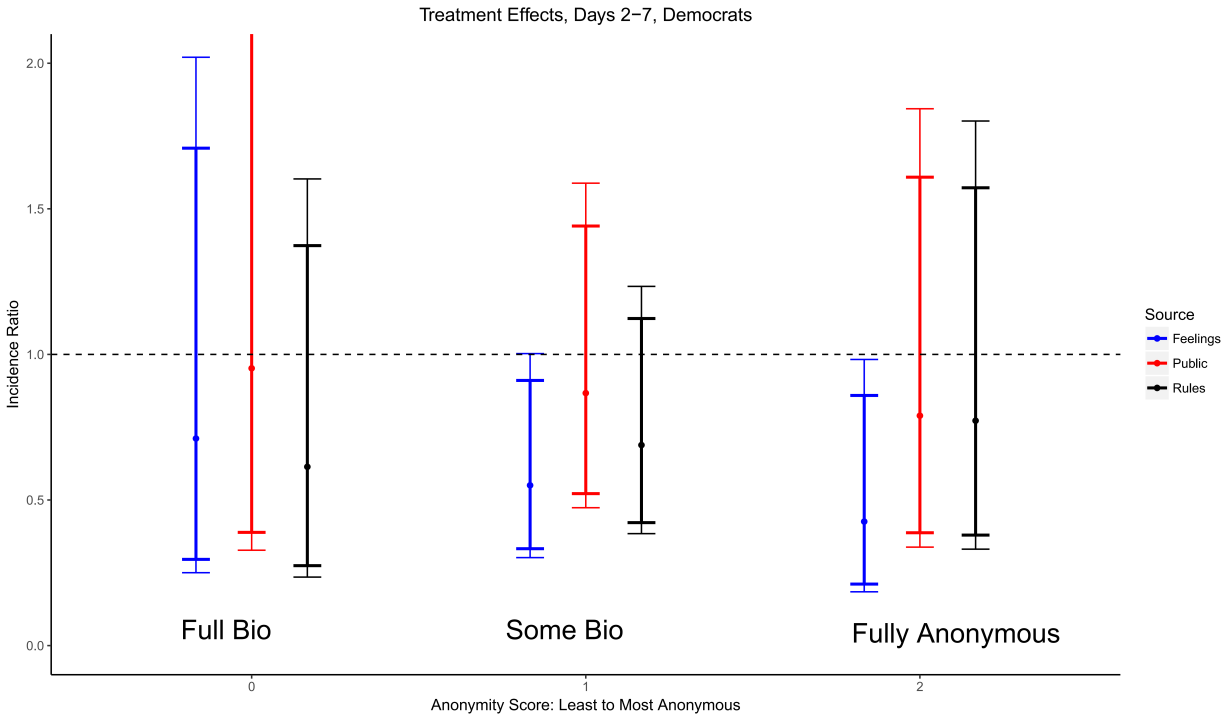


Figure 5: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week (excluding day 1; top panel) and third and fourth weeks (bottom panel) after treatment. For example, the Incidence Ratio associated with the Feelings treatment on subjects with Anonymity Score 1 in the middle of the top panel means that these subjects sent 51% as many directed incivil 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.

Change in Incivility Among Democrats (*N=147*)



Change in Incivility Among Republicans (*N=163*)

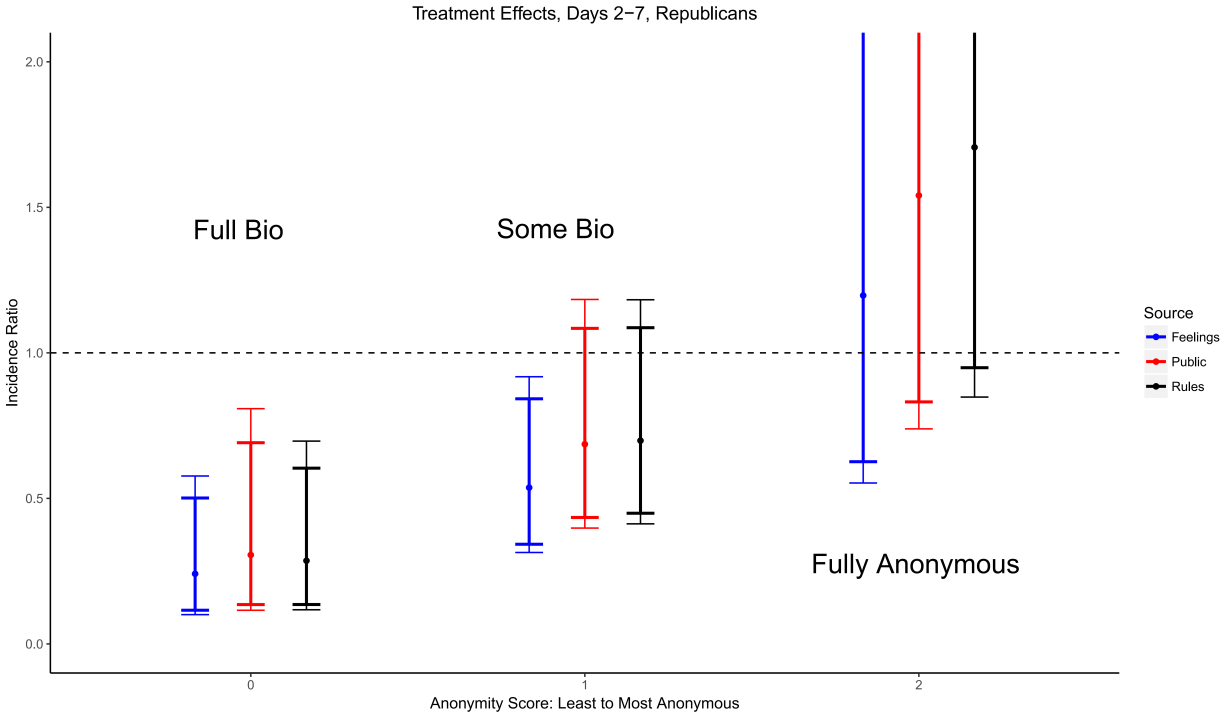
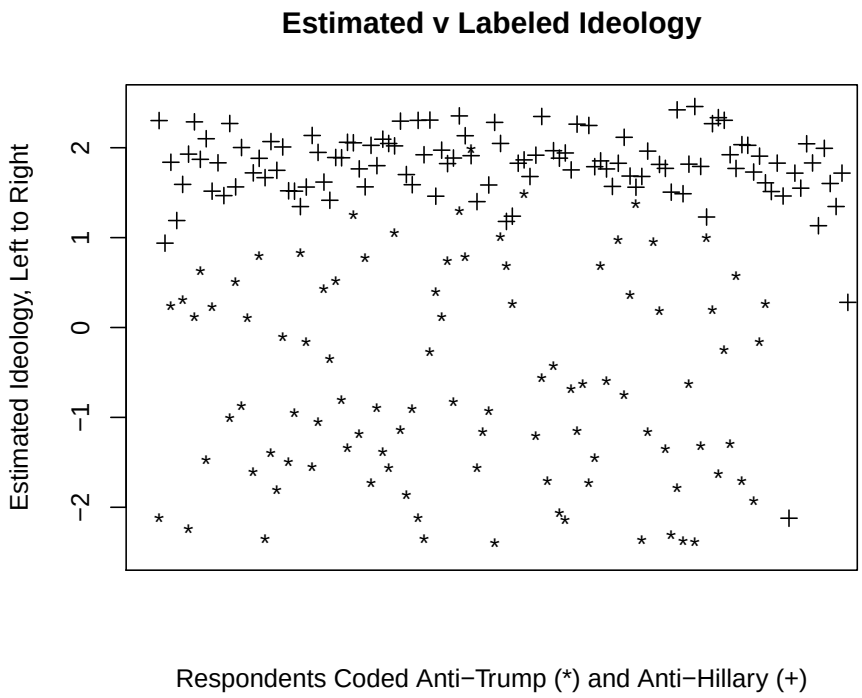


Figure 6: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week (excluding day 1) after treatment, divided by subject ideology. For example, the Incidence Ratio associated with the Feelings treatment on subjects with Anonymity Score 1 in the middle of the top panel (Democrats) means that these subjects sent 56% as many directed incivil 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.

Figure 7



The reaction of these fully anonymous Republican subjects is consistent with the presence of dedicated bad actors ("trolls") whose aim was to spread discord. During the campaign, Hillary Clinton's campaign website published an article explaining how "alt-right" trolls were using anonymous Twitter accounts and were being retweeted by Donald Trump (Chan, 2016). The article identified "Pepe the Frog" as a symbol of this group, and indeed, many of the anonymous Republicans in my sample had an image of Pepe as their Twitter bio photo. That the Rules treatment had the largest positive effect is unsurprising: telling people who were intentionally antagonizing others for fun that they were breaking the "rules of political civility" was tantamount to a congratulations.

One possible explanation for the lack of an effect on Democrat subjects is that this group was more heterogeneous. I implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points. As Figure 7 demonstrates, there was significant heterogeneity in the ideal points of subjects I coded as Democrats, but not for Republicans.

All but two of the subjects coded as Anti-Hillary (Republicans) had estimated ide-

ology scores above 1, and only one was coded as left of center. However, a full third of the subjects coded as Anti-Trump (Democrats) 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 seems that there was a significant contingent of moderate Anti-Trump Republicans that I classified as Democrats. Because the Feelings and Rules treatment messages were explicitly designed to appeal to subjects' partisan group identities (and identified the Anti-Trump subjects as "Democrats"), the ideological heterogeneity within this group could pose a problem for estimating average treatment effects.

If I restrict the analysis of Democrats in Figure 6 to only those with estimated ideology scores to the left of center, I find support for this *ex post* explanation. The point estimates for the two moral suasion treatment effects become more negative, seen in Figure 8. Because the sample size is down to 86, the previously significant effects of the Feelings treatment are no longer significant, but the largest change is on the Rules effects, which are now significantly negative for all but the fully anonymous subjects.

In keeping with the claim that the weak findings on the full Democrat sample is that it actually contained Republicans, the results for the Public treatment are essentially unchanged between Figures 6 and 8: unlike the other two treatments, this message did not refer to its recipients as "Democrats." As such, there was no possibility of partisan misidentification.

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 liberals 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 more effective at causing Republicans to decrease their rate of incivil tweeting. The difference in the

Change in Incivility Among "Real" Democrats (N=86)

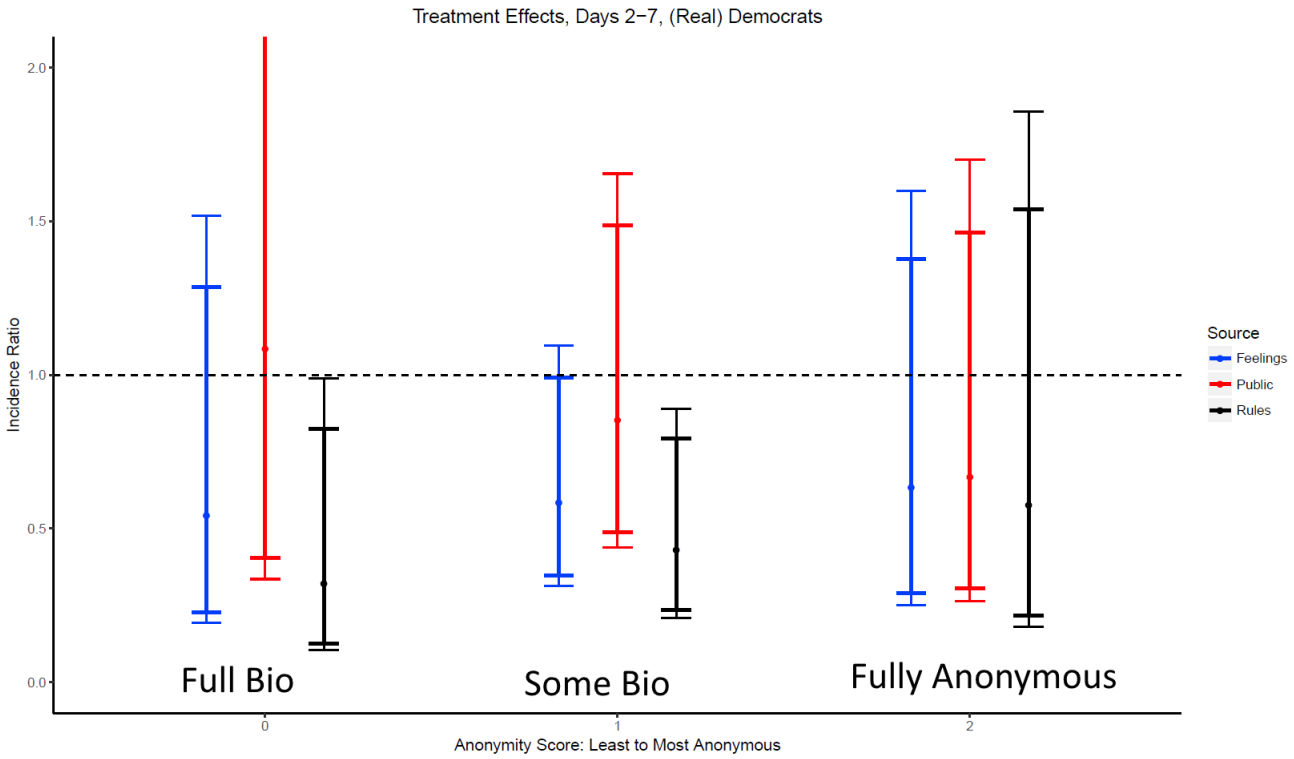


Figure 8: The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week (excluding day 1) after treatment, restricted to Anti-Trump subjects with an estimated ideology left of center. For example, the Incidence Ratio associated with the Feelings treatment on subjects with Anonymity Score 1 in the middle of the top panel (Democrats) means that these subjects sent 55% as many directed incivil 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.

effects on Republicans and Democrats becomes smaller when I remove subjects classified as Democrats who may actually have been right of center, but a gap remains.

I failed to find support for Hypothesis 1: there was little difference between sanctioning language designed to appeal to subjects' moral sense of Care or Authority. My *post-hoc* explanation for this conclusion is that this election was not normal. Following Haidt (2012), I expected that a message reminding subjects of the rules of political civility would be more effective on Republicans, but in the 2016 US Presidential election, it was Democrat Hillary Clinton who explicitly positioned herself on the side of civility.

Further, the lack of a response from Democrats to the Feelings treatment may be explained by the tweets they sent to my bots in response to being sanctioned. In several cases, Democrats told my bots something like "these other people are Trump supporters, so I don't care about their feelings"; no Republicans expressed a similar sentiment. The Trump campaign elicited extremely strong reactions from some Democrats, so it is possible that this resistance to moral suasion based on Care was idiosyncratic to the 2016 election. Although the uniqueness of the campaign may explain my unexpected findings, it is difficult to read these results as support for Haidt's model.

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 incivil tweets

at partisans from the other side, is 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, especially Republicans.

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 incivil language on forums where others have already been incivil. 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 incivil 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. The stakes of improving online political discourse are high: the social web could fulfill the promise of widespread deliberative democracy. If partisan incivility becomes further established as the norm in online communication, it could lead to further affect polarization and self-segregation, creating entirely separate epistemic communities and rendering deliberation impossible.

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 incivil 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 the 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 incivil 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 1: Attrition Rates and Causes

	Control	Democrats	Republicans
Initial assignment	108	104	118
Failed treatment application	0	2	2
Tweeted too often/bots	3	1	5
Suspended	0	1	2
Weird	2	0	0
Final	102	100	108
Attrition	6%	4%	8%

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 OLS Specification of Main Results

The dependent variable of interest in this analysis is the number of times a subject sent an incivil 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 the dependent variable.

The results in Figure 9 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 9 shows the same analysis but with the 61 misclassified Democrats (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 e negative binomial regression is the correct model, but researchers might reasonably disagree and assign less credibility to the persistence of the effects.

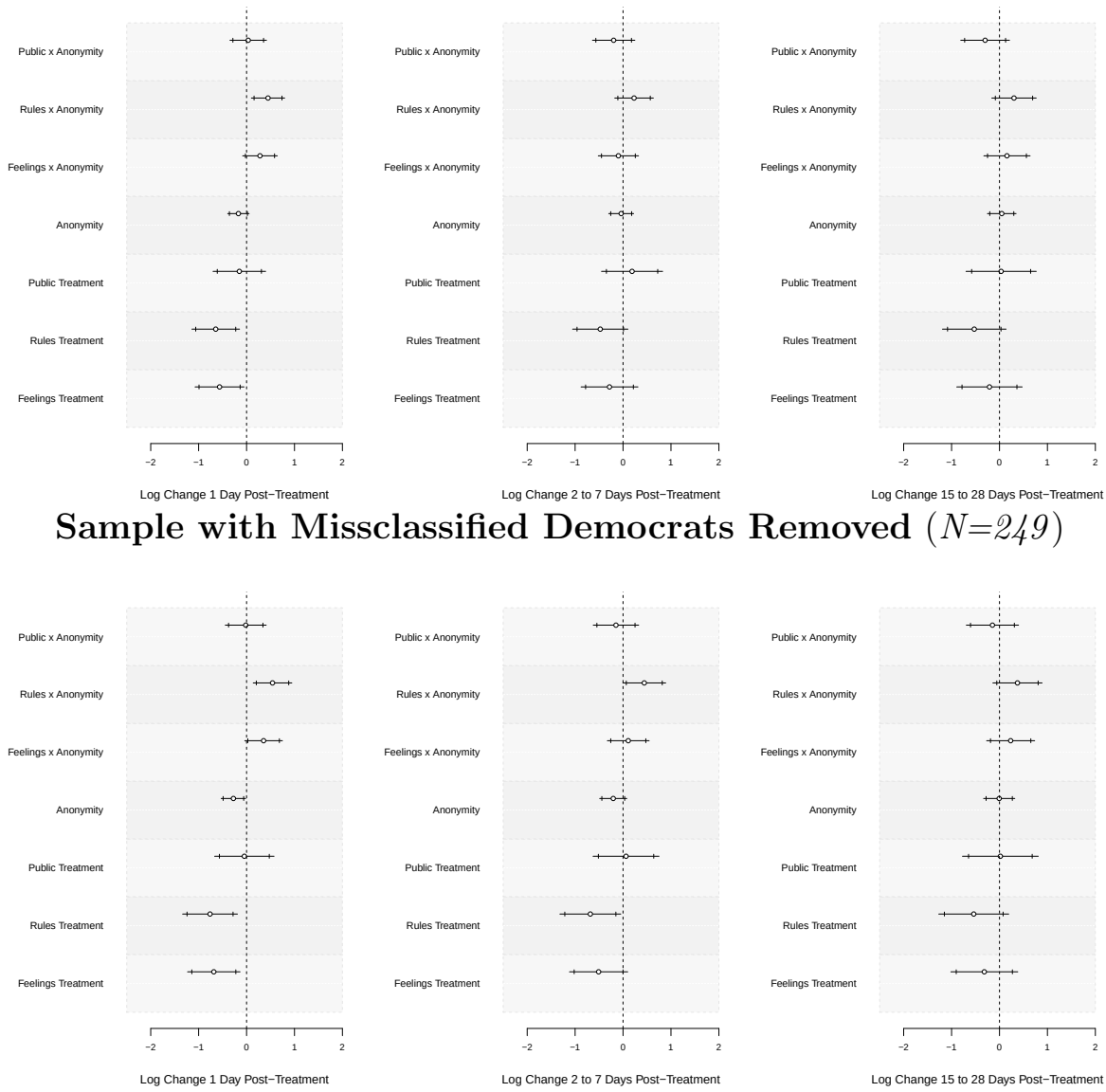


Figure 9: Each panel represents the results of a separate OLS regression in which the outcome variable is the log of the number of times a subject directed an incivil tweet at another user in the specified time period. The top three plots are calculated only on the Liberal sample, and the bottom three 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

Barberá, Pablo. 2014. "How social media reduces mass political polarization. Evidence from Germany, Spain, and the US." *Job Market Paper, New York University* .

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91.

Barberá, Pablo, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker, and Sandra González-Bailón. 2015. "The critical periphery in the growth of social protests." *PloS one* 10 (11): e0143611.

Barberá, Pablo, Richard Bonneau, Patrick Egan, John T Jost, Jonathan Nagler, and Joshua Tucker. 2014. Leaders or followers? Measuring political responsiveness in the US Congress using social media data. In *110th American Political Science Association Annual Meeting*.

Bejan, Teresa M. 2017. *Mere Civility*. Harvard University Press.

Berry, Jeffrey M, and Sarah Sobieraj. 2013. *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

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

Buckels, Erin E, Paul D Trapnell, and Delroy L Paulhus. 2014. "Trolls just want to have fun." *Personality and individual Differences* 67: 97–102.

Chan, Elizabeth. 2016. "Donald Trump, Pepe the frog, and white supremacists: an explainer.".

Chen, Adrian. 2015. "The Agency." *New York Times Magazine* June 2, 2015.

Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions.".

Duggan, M, and A Smith. 2016. "The political environment on social media." *Pew Research Center* 25.

Earl, Jennifer, Heather McKee Hurwitz, Analicia Mejia Mesinas, Margaret Tolan, and Ashley Arlotti. 2013. "This protest will be tweeted: Twitter and protest policing during the Pittsburgh G20." *Information, Communication & Society* 16 (4): 459–478.

Fishkin, James S. 2011. *When the people speak: Deliberative democracy and public consultation*. Oxford University Press.

Frijda, Nico H. 1988. "The laws of emotion." *American psychologist* 43 (5): 349.

Greenwood, S, A Perrin, and M Duggan. 2016. "Social Media Update 2016." *Washington, DC: Pew Internet & American Life Project. Retrieved November* 27: 2016.

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* .

Haidt, Jonathan. 2001. "The emotional dog and its rational tail: a social intuitionist approach to moral judgment." *Psychological review* 108 (4): 814.

Haidt, Jonathan. 2012. "The righteous mind: Why good people are divided by politics and religion.".

Hilbe, Joseph M. 2008. "Brief overview on interpreting count model risk ratios: An addendum to negative binomial regression.".

Hindman, Matthew. 2008. *The myth of digital democracy*. Princeton University Press.

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* .

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*.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes. 2012. "Affect, not ideology a social identity perspective on polarization." *Public opinion quarterly* 76 (3): 405–431.

Iyengar, Shanto, and Sean J Westwood. 2015. "Fear and loathing across party lines: New evidence on group polarization." *American Journal of Political Science* 59 (3): 690–707.

Karlsen, Rune, and Eli Skogerbø. 2013. "Candidate campaigning in parliamentary systems Individualized vs. localized campaigning." *Party Politics* p. 1354068813487103.

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

Ladd, Jonathan M. 2011. *Why Americans hate the media and how it matters*. Princeton University Press.

Lelkes, Yphtach, Gaurav Sood, and Shanto Iyengar. 2015. "The hostile audience: The effect of access to broadband Internet on partisan affect." *American Journal of Political Science* .

Milner, Ryan M. 2013. "FCJ-156 Hacking the Social: Internet Memes, Identity Antagonism, and the Logic of Lulz." *The Fibreculture Journal* (22 2013: Trolls and The Negative Space of the Internet).

Munger, Kevin. 2016. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* pp. 1–21.

Munger, Kevin, Patrick Egan, Jonathan Nagler, Jonathan Ronen, and Joshua A Tucker. 2016. "Learning (and Unlearning) from the Media and Political Parties: Evidence from the 2015 UK Election.".

Mutz, Diana C. 2015. *In-your-face politics: The consequences of uncivil media*. Princeton University Press.

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.

O'Reilly, Bill. 2003. *The no spin zone: Confrontations with the powerful and famous in America*. Three Rivers Press.

Papacharissi, Zizi. 2002. "The virtual sphere The internet as a public sphere." *New media & society* 4 (1): 9–27.

Papacharissi, Zizi. 2004. "Democracy online: Civility, politeness, and the democratic potential of online political discussion groups." *New Media & Society* 6 (2): 259–283.

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

Settle, Jaime. Forthcoming. *Newspaper to News Feed: How the Social Communication of Politics Affectively Polarizes the American Public*.

Theocharis, Yannis, Pablo Barberá, Zoltan Fazekas, and Sebastian Adrian Popa. 2015. "A Bad Workman Blames His Tweets? The Consequences of Citizens Uncivil Twitter Use When Interacting with Party Candidates." *The Consequences of Citizens Uncivil Twitter Use When Interacting with Party Candidates (September 5, 2015)* .

Trippi, Joe. 2004. "The revolution will not be televised." *CAMPAIGNS AND ELECTIONS* 25 (8): 44–44.

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

Wojcieszak, Magdalena E, and Diana C Mutz. 2009. "Online groups and political discourse: Do online discussion spaces facilitate exposure to political disagreement?" *Journal of communication* 59 (1): 40–56.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2016. "Ex Machina: Personal Attacks Seen at Scale." *arXiv preprint arXiv:1610.08914* .