Experimentally Reducing Partisan Incivility on Twitter

Kevin Munger

New York University, contact: km2713@nyu.edu

September 7, 2017

Abstract

Cross-partisan incivility is a pressing concern in American politics, especially online. When partisans communicate incivilly, they are less likely to learn from each other and more likely to distrust each other. This paper analyzes the way that people learn about norms of political behavior online. I conduct an experiment that tests how different forms of moral suasion—appeals to morality intended to influence behavior—affect how Democrats and Republicans learn about norms of partisan speech. Using bots that shared the political identity of the subjects, I sent messages that appealed to the moral principle theorized to be most convincing to either liberals ("care") or to conservatives ("authority"). Using a sample of subjects who had been frequently incivil in political discussions on Twitter during the 2016 US presidential election, I found that both forms of moral suasion were equally effective at dissuading both Democrats and Republicans from online incivility. These effects were significantly moderated by the anonymity of the subjects, especially among Republicans: subjects who elected to have an anonymous profile were much less likely to change their behavior. On

some subsamples, the reduction in incivility persisted for up to a month after treatment.

1 Introduction

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

The trend towards incivility in 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 and reduces the gatekeeping power of traditional actors; the invention of the printing press led to elite concern about civil discourse during the time of the Reformation (Bejan, 2017). Norms of political discourse on the internet govern and are governed by the behavior of millions of Americans, an unprecedented feature that represents the massive erosion of gatekeeping power once held by broadcast media.

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 between Democrats and Republicans (Iyengar, Sood, and Lelkes, 2012). Affect polarization is directly related to the decline of 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 are not solely responsible for 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.

Partisan incivility is increasingly the norm in online interactions, but norms can be

changed. The goal of the current study is to test the theory that people update their beliefs about the proper norms of behavior based on arguments that appeal to the moral foundations of their beliefs. I conducted an experiment that tests different interventions aimed at promoting civil political discourse during the 2016 US Presidential election. Extending the method I pioneered in Munger (2017), I used Twitter accounts that I created and controlled ("bots") to send messages chastising users engaged in incivil cross-partisan 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 over a continuous and unbounded time frame.

Users were sampled by searching for tweets that mentioned either Donald Trump or Hillary Clinton and which were directed at another, non-elite user. I used an algorithm to select the most incivil of these tweets. I randomly assigned the subject to a treatment arm and used bots[1] to send them a message.

The messages were only sent by bots that shared the partisan identity of the subjects (eg Republican bots messaged Republican subjects) because social norms are more easily spread among people with a shared social identity. The primary experimental manipulation was to vary the language of the messages sent to subjects. I sent messages that appealed to the moral principle theorized to be most convincing to either liberals ("care") or to conservatives ("authority"), as well as another message with no moral appeal. I also kept a group of subjects as a true control group, and did not send them any message.[2]

I found that both forms of moral suasion were equally effective at dissuading both Democrats and Republicans from online incivility, and that the non-moral message was only slightly less effective. While it is encouraging that I have demonstrated a realistic example of how norms of political behavior promulgate, the symmetry between Republicans and Democrats does not support my hypothesis that the effectiveness of moral appeals varies with the ideology of the subject. This null finding lends support to the theory that the shared social identity of the bots and the subjects is the primary mechanism by which people adopt new norms of political behavior. This is in accordance with my previous finding in Munger (2017) that white men were more likely to accept

[1]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.

[2]The research design, dependent variable measurement, and main hypothesis were pre-registered at EGAP.org (number 20160921AA) prior to any research activities.

a norm against using anti-black slurs when that norm was promoted by a white bot than when it was promoted by a black bot.

As predicted in this previous experiment, treatment effects were significantly moderated by the anonymity of the subjects: subjects who provided less personal information on their Twitter profiles were less amenable to normative pressure. This effect heterogeneity was especially pronounced among Republicans, to the point that all three messages caused a (nonsignificant) reaction *against* the norm being promoted and increased the use of incivility among anonymous Republicans. I believe that many of these subjects may have been "trolls," tweeting in bad faith with the explicit aim of upsetting others and disrupting civil discourse.

These results contribute to a growing literature on the importance of social identity in determining how the norms of political behavior are spread. The example of cross-partisan incivility represents something of a hard case for this kind of norm promotion: people engaged in this behavior are angry and defensive (Iyengar and Westwood, 2015), but a single message from a stranger who shared their partisan identity caused them to change their behavior. This example is also substantively important; incivil cross-partisan communication fails to bring about the moderation that can occur when people talk through their differences, but instead drives people further apart. 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). If online incivility is the dominant norm of cross-partisan communication, deliberative democracy cannot take place.

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 over 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)).

The implications of this democratization were not fully understood at the time. One important consideration is that 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)" (Hindman, 2008).

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 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 people 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 facilitated by 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 online speech norms. Today, the internet is widely

regarded as rife with offensive or harassing speech designed to mock sincere expression—incivility is dominant online (Buckels, Trapnell, and Paulhus, 2014; Milner, 2013).

The extent to which incivility obtains depends on the specific technical affordances of different online platforms. The most important feature 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...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 increasing animosity between Democratic and Republican partisans. Scholars have described this animosity as "affect polarization." Partisans dislike each other (Iyengar, Sood, and Lelkes, 2012) and tend to trust co-partisans and distrust out-partisans (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, 2017).

The uptick in partisan polarization began well before the mass adoption of social media, and scholars remain divided as to whether social media causes increased polarization (Barberá, 2014; Lelkes, Sood, and Iyengar, 2015; Settle, Forthcoming). 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. More importantly for the mass public, this behavior means that citizens who wish to engage with politicians or each other in response to a politician's 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 engage in public discussions. There is also 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 established incivility as normal in 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". 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; James 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 causes 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. I conducted an experiment to test the mechanisms underlying online norm promotion aimed at decreasing cross-partisan incivility.

The first step in performing this experiment was finding conversations that were incivil, between out-partisans, *and* about politics. In early October 2016, I used the streamR package to scrape Twitter in real time for tweets mentioning either "@realDonaldTrump" or "@HillaryClinton"—the Twitter accounts of the two major party candidates in the 2016 US presidential election. I only kept tweets that were directed at another user who was *not* either Trump or Clinton.

In this way, I found a sample of tweets from non-elites that were concerned with the topics 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

Figure 1: Finding Non-Elite Incivility

Dixon (2017) 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 an out-partisan. Many of the potential subjects I found this way were tweeting at elites—either people who were "verified" (Twitter's method of indicating public figures), 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.

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.[3] Ty then responded to Parker 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 Republican,

[3]I censor the usernames of the subjects to preserve their anonymity.

and sent him a tweet from one of my Republican-identified bots. For a visual overview of this selection process, see Figure 2.

Based 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!", 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; see Figure 3.

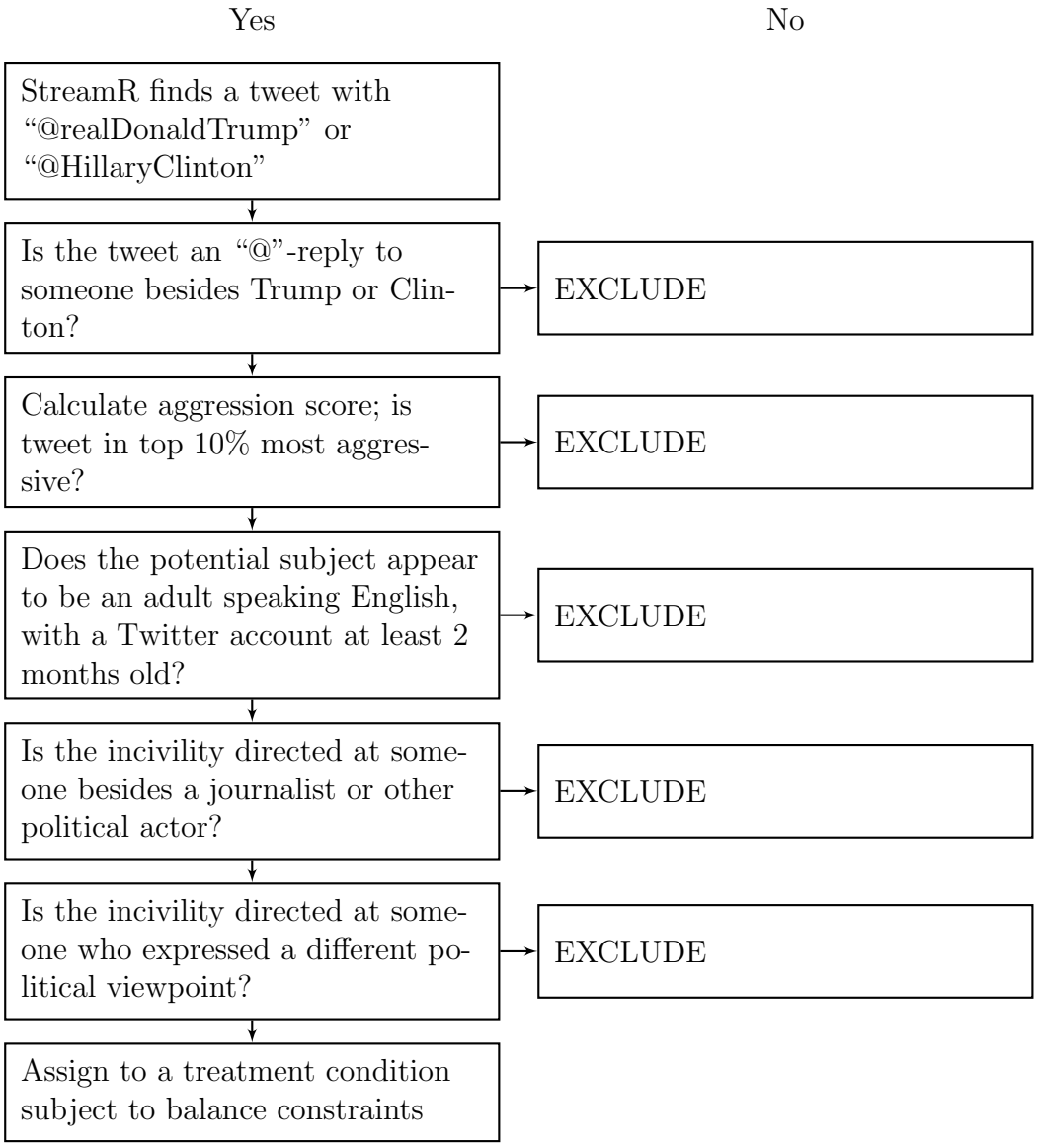
I created four bots. Neil, in panel (a), was a bot who appeared to be pro-Clinton. The other three were pro-Democrats, pro-Trump, and pro-Republicans (see Todd, in panel (b)). To manipulate these identities, I changed the large banner in the middle of the profile, the small logo in the bottom right of the bots' profile pictures, and the "bio" field below their usernames (eg "Hillary 2016!"; "Republicans 2016!"). The four bots were otherwise identical. All of the bots appeared to be white men, keeping the 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. Munger (2017) varied the number of followers that sanctioning bots had, and found that bots with few followers had very little effect. Based on this finding, I purchased around 900 followers for each of my four bots. 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.

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 ("Care", "Authority", or "Public") sent by one of two bots (pro-candidate or pro-

Figure 2: Sample Selection Process



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

Figure 3: (a) Example Bot—Clinton Condition

(b) Example Bot—Republican Condition

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.[4] 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.

After I tweeted at a subject, he or she 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 and see the bots' partisan leaning.

The primary outcome of interest was how subjects changed their behavior after having been sanctioned. This measure was computed by comparing the change in the rate at which each subject sent incivil tweets before and after being sanctioned to the change in that rate for a control subject who received no message.

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 is *in-group social norm promotion*: to cause subjects to update their beliefs about correct normative behavior for someone sharing their social identity. Munger (2017) finds that sanctions 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 efficacy of moral suasion. I based my approach on the moral intuitionist model proposed by Haidt (2001), who 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 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

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

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 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 place significant weight on all six of these moral foundations. Westerners on the left of the political spectrum, however, 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 moral 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.

Effective moral suasion appeals to the correct moral foundation of the subject. To that end, I designed two different treatments. The first appealed to the Care moral foundation, and I thus expected it 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 treatments, I included a non-moral 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. I used a message that would serve to remind subjects that their incivil tweets were public. My hypothesis was that this treatment would decrease the subjects' use of incivility, but that the effect would be smaller than the moral treatments. The message I used 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 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 open these users are to messages from co-partisans promoting norms of civility; more anonymous people are less invested in their identities, and thus less amenable to changing their behavior. [5]

Hypothesis 2 *The reduction in incivility caused by the treatments will be larger for less anonymous subjects.*

5 Results

The behavioral outcome in this experiment is partisan incivility targeted at other Twitter users. 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 again used the model trained by Wulczyn, Thain, and Dixon (2017) to assign an "aggression score" (between 0 and 1) to each of these 367,000 tweets. This measure was skewed toward the lower end of the distribution, so I selected all tweets above the 75th

[5]Note that this hypothesis was recorded in the pre-analysis plan for my previous experiment, EGAP registration number 20150520AA. In Munger (2017) I explain how this hypothesis was not supported in this other context: in a sample of users who were using racist slurs to harass others, I found that more anonymous users were actually *more* amenable to normative pressure. I argue that this counter-intuitive finding was due to the peculiarity of the sample: people who were already tweeting the word "n****r" from an account that included their name and a picture of themselves were likely committed racists or white supremacists.

percentile aggression score and coded them as incivil.[6]

To control for each subject's 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 (Hilbe, 2008).[7] The negative binomial specification is estimated using the following model:

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

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

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 (calculated as the Incidence Ratio, IRR) of the Care treatment on subjects with Anonymity Score 1 (the middle category) is:

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

Note that the IRR is a ratio: 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 appears longer than the lower half.

The experimental results on the full sample without interaction effects for subject anonymity are displayed in Figure 4; in all of the analysis that follows, the dependent variable is the change from the subjects' pre-treatment rates of sending incivil tweets to their post-treatment rates of sending incivil tweets, relative to a subject in the control condition. The Care treatment caused a significant reduction in the first day after treatment, and both the Care and Authority treatments caused a significant reducition in the first week after treatment. Note that these time periods are non-overlapping;

[6]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.

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

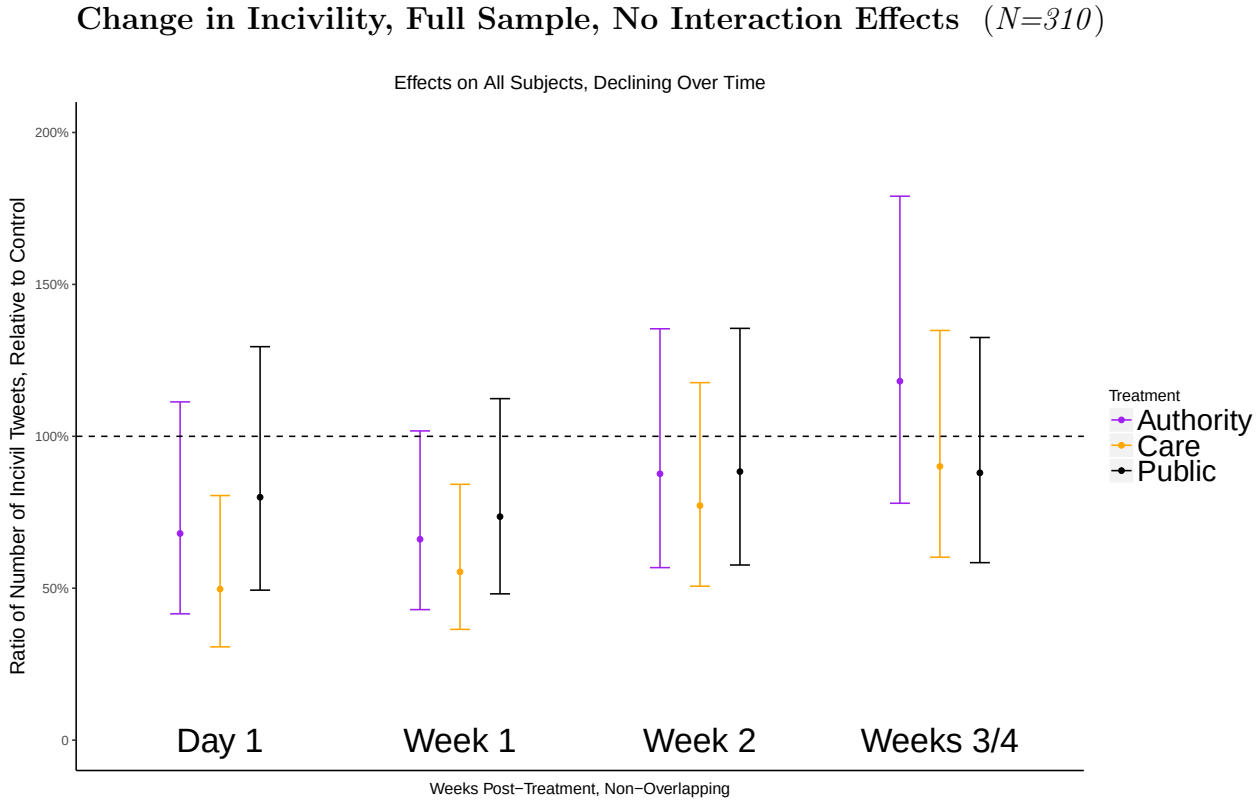


Figure 4: The Incidence Ratio calculated from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 50% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

"Week 1" consists of days 2-7, and "Week 2" of days 8-14. As expected, the Public treatment condition had an effect in the same direction, but it was smaller than the effect of the two moral treatments. Moving from left to right in Figure 4, we see the treatment effects decay over time.

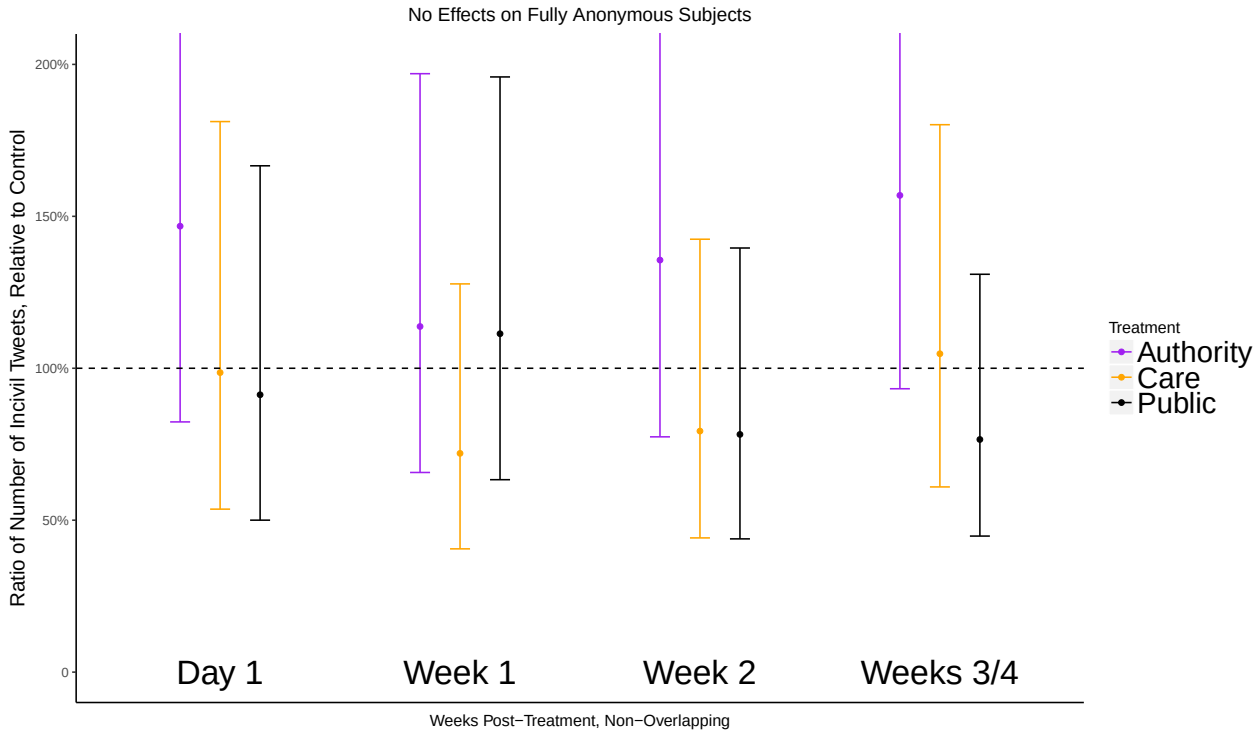
$IRR_{care} = 0.50$ in Day 1 can be seen in the orange line on the left of the plot. This Incidence Ratio means that the average subject who received the Care treatment sent 50% as many directed incivil tweets as the average subject in the control condition.[8]

The baseline model presented in Figure 4 does not capture the predicted effect heterogeneity based on subject anonymity. Figures 5, 6, and 7 show the model results with an interaction term, and plot the effects on the fully anonymous, partially anonymous, and non-anonymous samples, respectively.[9]

[8]Note that this approach assumes that treatment effects are constant, and holds the pre-treatment level of aggressive tweets constant at its mean level.

[9]The confidence intervals in these figures are calculated from the variance of the following estimator:

Figure 5: **Change in Incivility, Anonymous Sample** (*N=133*)



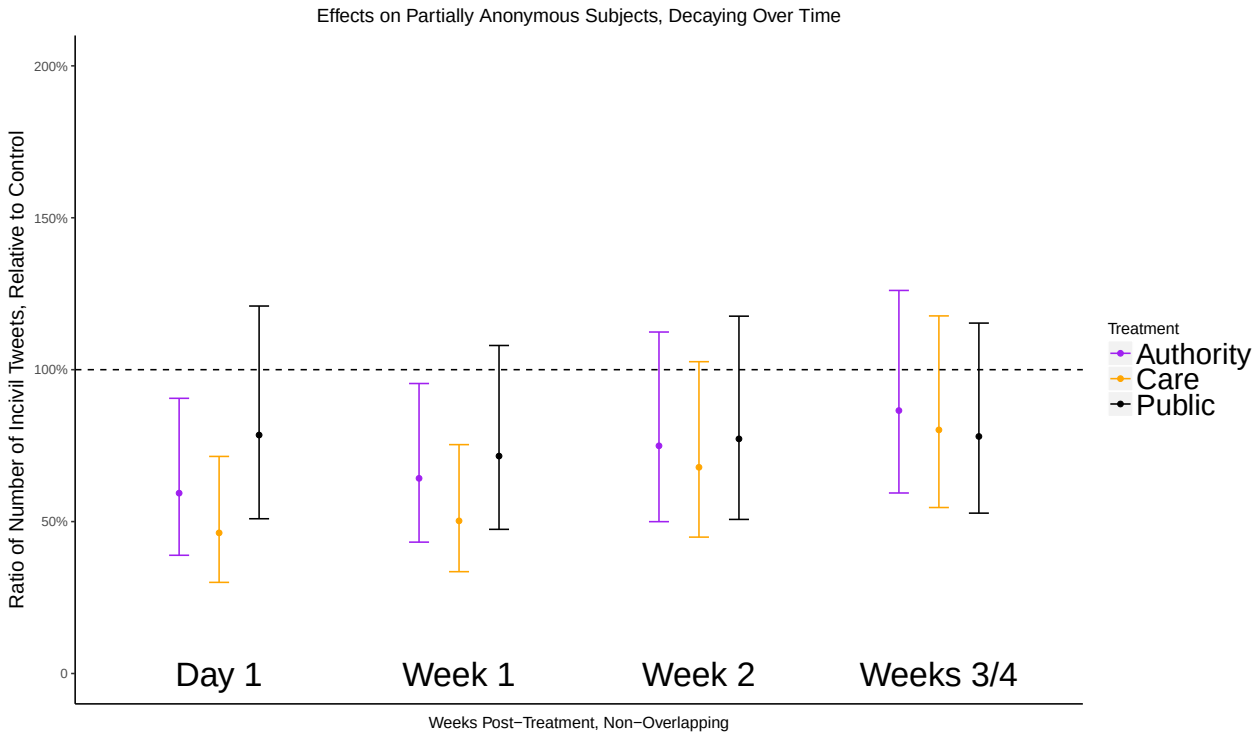
The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 98% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

Figure 5 shows that anonymous subjects did not respond to the treatment. Although none of the effects are significant, the point estimate on the Authority treatment is actually *positive*. Figure 6, on the other hand, shows a significant reduction in incivility from semi-anonymous subjects who received either the Care or Authority message. Like in the full sample, these effects decay over time. The effects on the non-anonymous sample, in Figure 7, are the largest. Again, both the Authority and Care conditions cause a significant reduction in incivility, with the effect of the Public condition still negative but smaller (though now significant, in Week 1).

Overall, treatment effects were larger among subjects who shared personal information on their profiles, as predicted in Hypothesis 2.

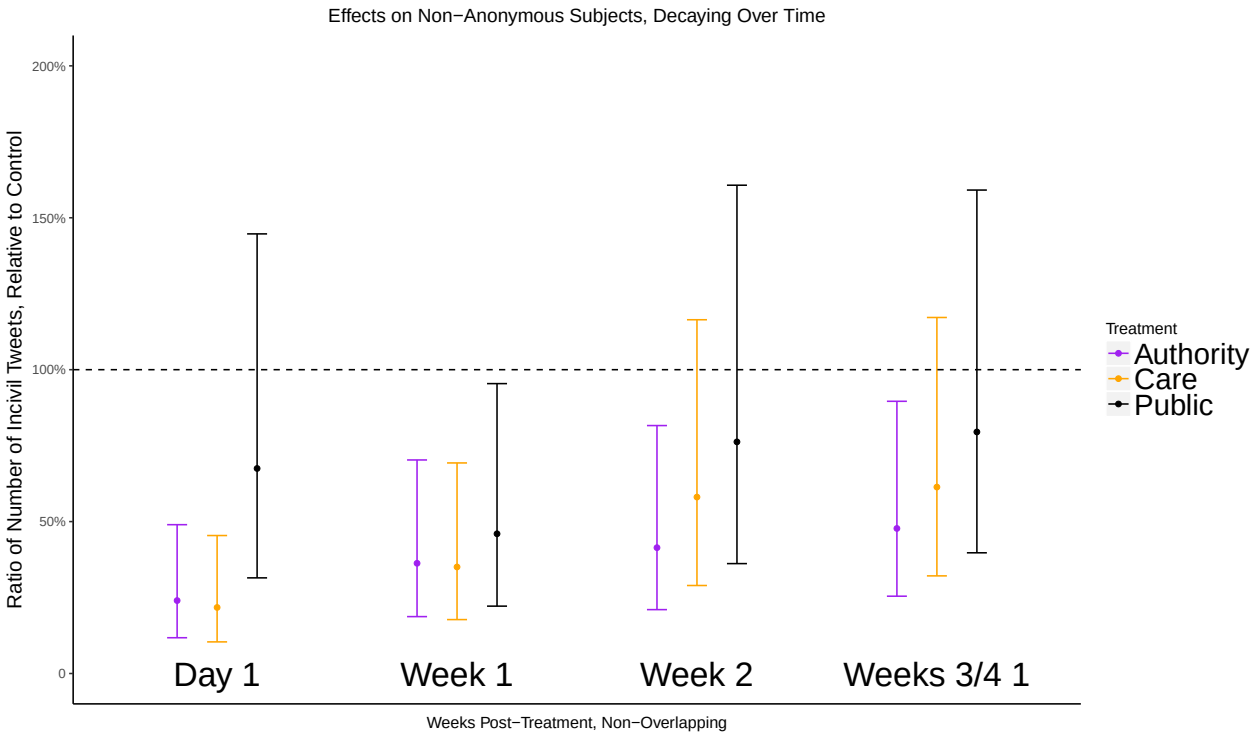
$$V_{feel \times Anon_1} = V(\hat{\beta}_2) + Anon^2 V(\hat{\beta}_6) + 2Anon \times Cov(\hat{\beta}_2 \hat{\beta}_6)$$

Figure 6: **Change in Incivility, Semi-Anonymous Sample** (*N=94*)



The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 47% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

Figure 7: **Change in Incivility, Non-Anonymous Sample** (*N=83*)



The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 30% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

To test Hypothesis 1, Figure 8 replicates the first analysis (without interaction effects), dividing the analysis between Republican (panel A) and Democrat (panel B) subjects. Contrary to my expectations, the effects of the different messages were essentially the same across both subject pools. In particular, the effects on Week 1 were identical: all three treatments caused a reduction in incivility, but only in the case of the Care condition was this reduction significant. One notable difference between the samples was in Day 1: Democrat behavior remained unchanged, while Republicans significantly reduced their use of incivility.

Dividing the subjects by *both* anonymity and partisan orientation entails a loss of statistical sample; the smallest subgroup has only 39 subjects. These models can be found in Appendix C. The most striking result is the degree to which anonymity moderates treatment effects on Republicans. The non-anonymous and partially anonymous Republicans react similarly to the full sample (and their reduction in incivility is in fact larger and more persistent than the full sample), but the effects on the fully anonymous Republicans are very different. The messages do not cause any reduction in incivility among this group, and the Authority treatment actually causes an *increase* in incivility.

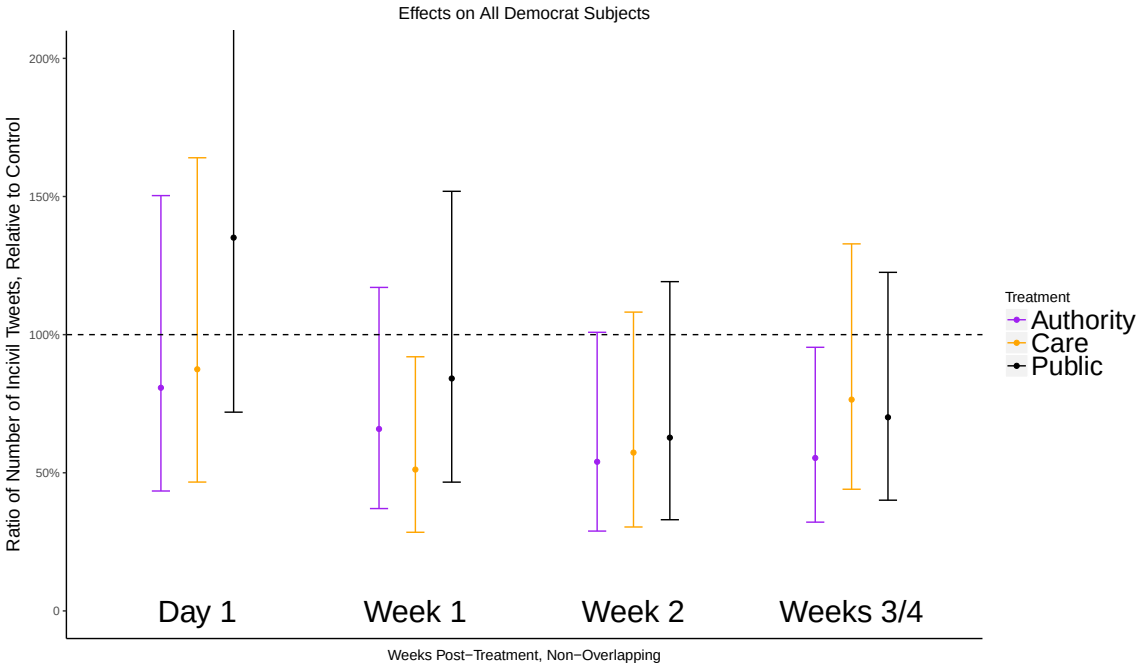
The reaction of these fully anonymous Republican subjects is consistent with the presence of dedicated bad actors ("trolls") whose aim was to spread discord. During the campaign, Hillary Clinton's campaign website published an article explaining how "alt-right" trolls were using anonymous Twitter accounts 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 Authority treatment had the effect of increasing incivility is in retrospect unsurprising: telling people who were intentionally antagonizing others for fun that they were breaking the "rules of political civility" was tantamount to a congratulation.

5.1 Estimating Subject Ideology Through Twitter Networks

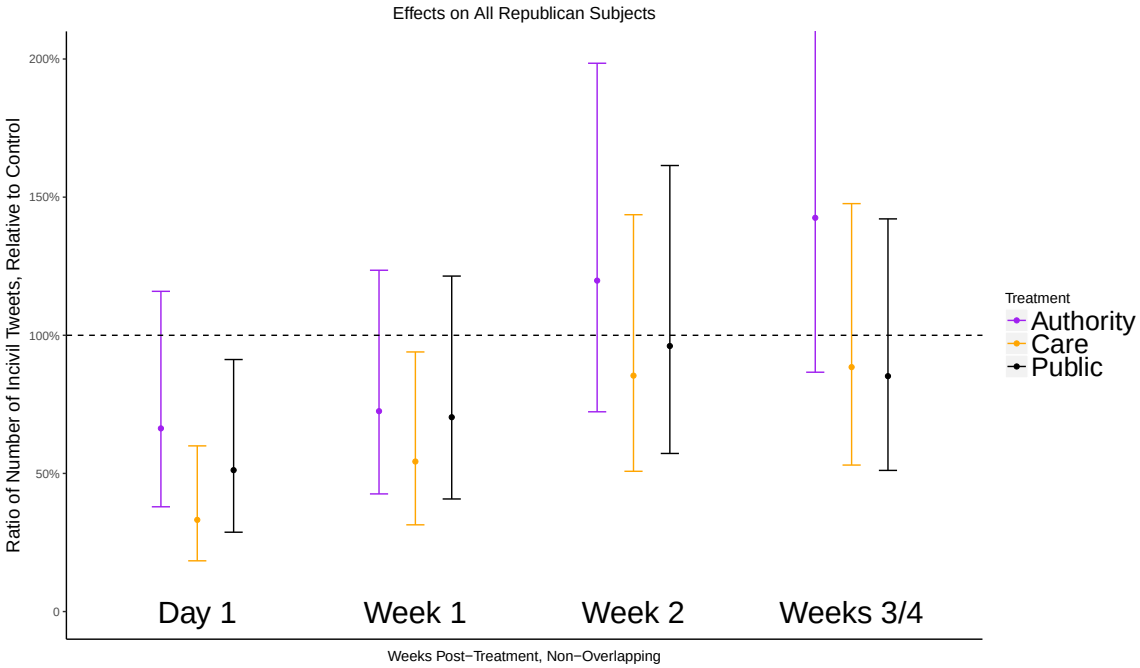
Overall, the point estimates of the effects of all three treatments on the sample of Democrats were in the predicted direction, but only the Care treatment was signficant, and only in the 1 Week time frame. This was not primarily due to heterogeneous effects based on subject anonymity; unlike the Republican sample, there is no consistent trend across the Democrat subsamples (see Appendix C).

The problem is that the "Democrat" sample is more ideologically heterogeneous—so

Figure 8: **Panel A: Change in Incivility Among Democrats** (*N=147*)

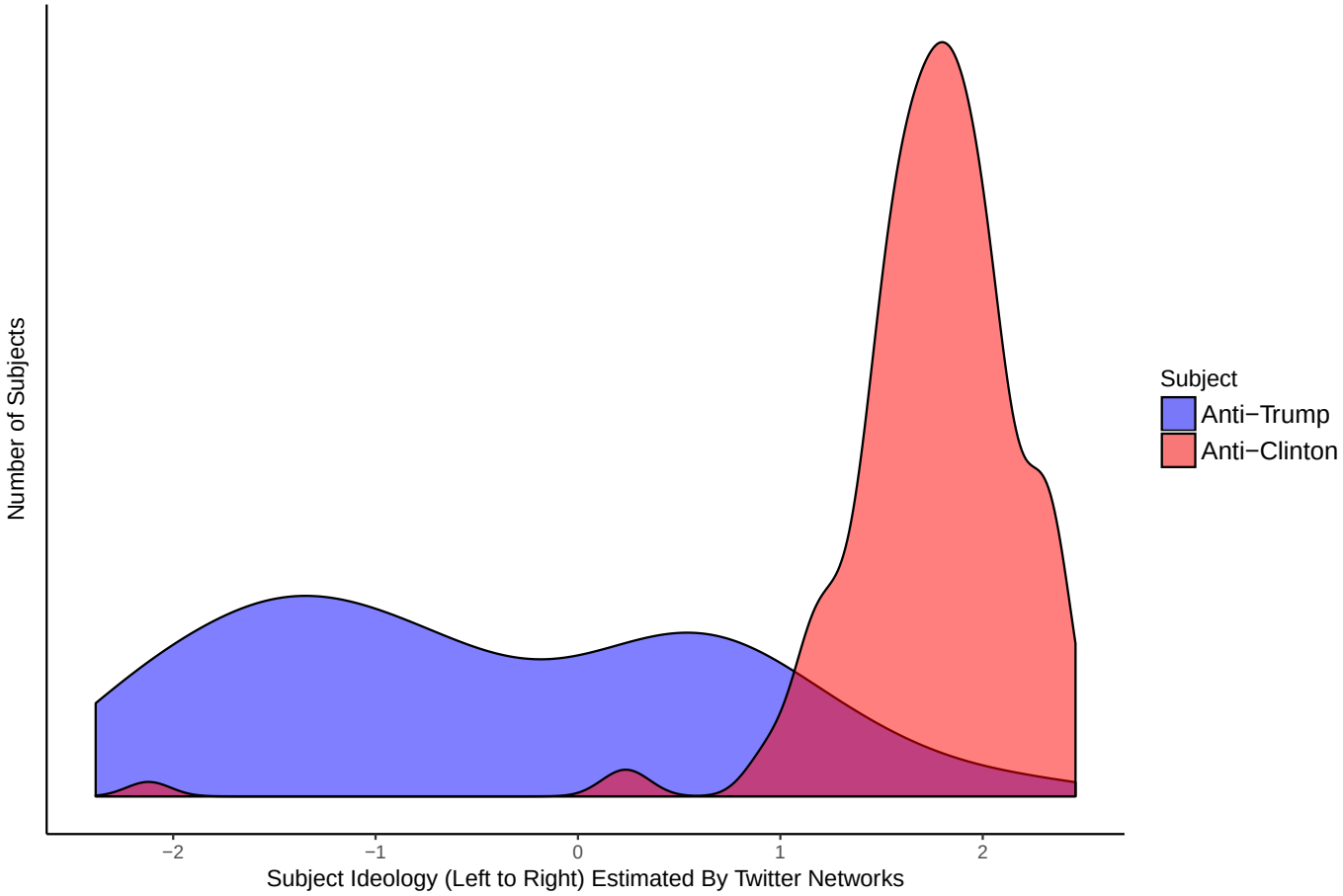


Panel B: Change in Incivility Among Republicans (*N=163*)



The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot in Panel A means that these subjects sent 90% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

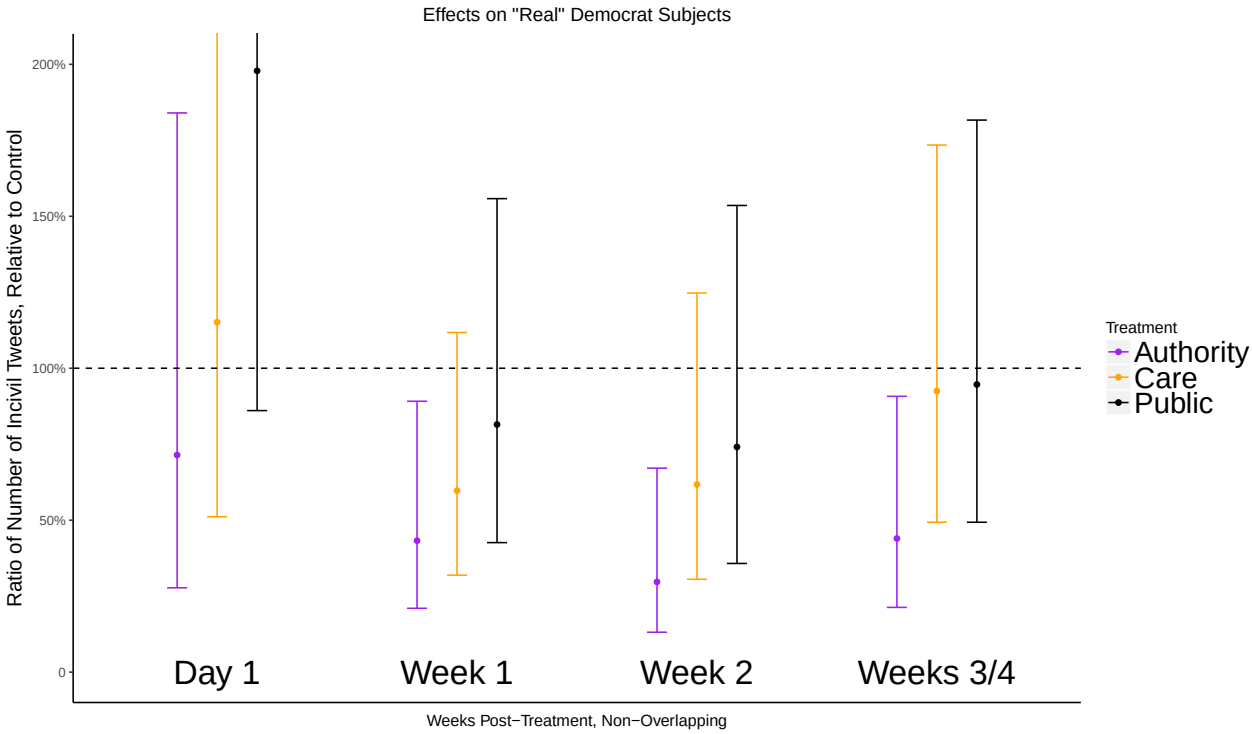
Figure 9: Anti-Trump Subjects were Ideologically Diverse



heterogeneous, in fact, that some of the subjects I coded as Democrats may actually have been Anti-Trump Republicans. I implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points based on the network of accounts they followed on Twitter. As Figure 9 demonstrates, there was significant heterogeneity in the ideal points of subjects I coded as Democrats (Anti-Trump), but not for Republicans (Anti-Clinton).

All but two of the subjects coded as Anti-Clinton (Republicans) 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 (Democrats) had estimated ideology scores right of center, although only a few were far to the right (have an ideology score above 1). Looking at Figure 9, the distribution of Anti-Trump subjects is bimodal; just to the right of the midpoint is a cluster of moderate Anti-Trump Republicans that I classified as Democrats. Because the Care and Authority treatment messages were

Figure 10: **Change in Incivility Among "Real" Democrats** (*N=86*)



The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 30% as many directed incivil tweets as the subjects in the control group. 95% confidence intervals.

explicitly designed to appeal to subjects' partisan group identities (and identified the Anti-Trump subjects as "Democrats"), the ideological heterogeneity within this group poses a problem for estimating average treatment effects.

If I restrict the analysis of Democrats in Figure 8 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 Authority treatment effect becomes more negative and significant, seen in Figure 10. Because the sample size is down to 86, the previously significant effects of the Care treatment are no longer significant at $p < .05$, but the point estimate is roughly the same as on the full sample.

In keeping with the claim that the weak findings on the full Democrat sample is because it actually contained Republicans, the results for the Public treatment are essentially unchanged between Figures 8 and 10: 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: during the previous election, 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."

This restraint does not extend to online political discussions. The technological affordances of social media make it far easier to be nasty, and incivil speech norms obtain. Civil cross-partisan communication might allow people to learn from and respect each other more, but when this communication is incivil, it has the opposite effect, driving partisans further apart and decreasing mutual trust and respect.

The experiment described in this paper studied the mechanism by which norms of political behavior are spread online in the context of cross-partisan incivility. I found that two kinds of moral appeals—one to the moral principle theorized to be most convincing to liberals ("care"), the other to conservatives ("authority")—were equally effective at changing the behavior of Republicans and Democrats, and somewhat more effective than a message with no moral content. The similar effectiveness of the two moral messages suggests that it was the shared partisan identity of the bots and the subjects that was responsible for causing subjects to change their behavior, rather than the language of the message they were sent.

An alternative (*post hoc*) explanation for the lack of a difference between language designed to appeal to subjects' moral sense of Care or Authority is that the 2016 election was idiosyncratic. 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. Indeed, on the sample of "real" Democrats with estimated ideology left of center, the "Authority" treatment caused a significant reduction in incivility on all but the fully anonymous subjects; I had theorized that this message would have no effect.

Further, the lack of a response from Democrats to the Care 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.

As expected, people who included more information on their Twitter profile were more responsive to all of the treatments. 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*.

The fully anonymous users in this sample may well have been intentionally using incivility as a strategic tool or as a source of morbid enjoyment; this makes them "trolls," distinct from the normal (if passionately polarized) people in the rest of the sample.[10] Although none of the treatment messages caused this subgroup to reduce their use of incivility, the Authority treatment actually caused an *increase* in incivility, especially among anonymous Republican subjects.

This finding concords with recent research on online trolling (Phillips and Milner, 2017), 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 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

[10]The term "troll" has referred to a variety of behaviors in the short history of the internet, and is now sufficiently capacious that its use risks confusion (Coleman, 2014; Phillips, 2015). I define a "troll" as someone who posts in bad faith: the content of what they write is meaningless except insofar as it accomplishes their goal of causing confusion or pain.

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 11 are very similar to those in Figure 4. The point estimate for the Authority treatment is largest, followed by the Care 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 11 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 Care and Authority 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 Care and Authority 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.

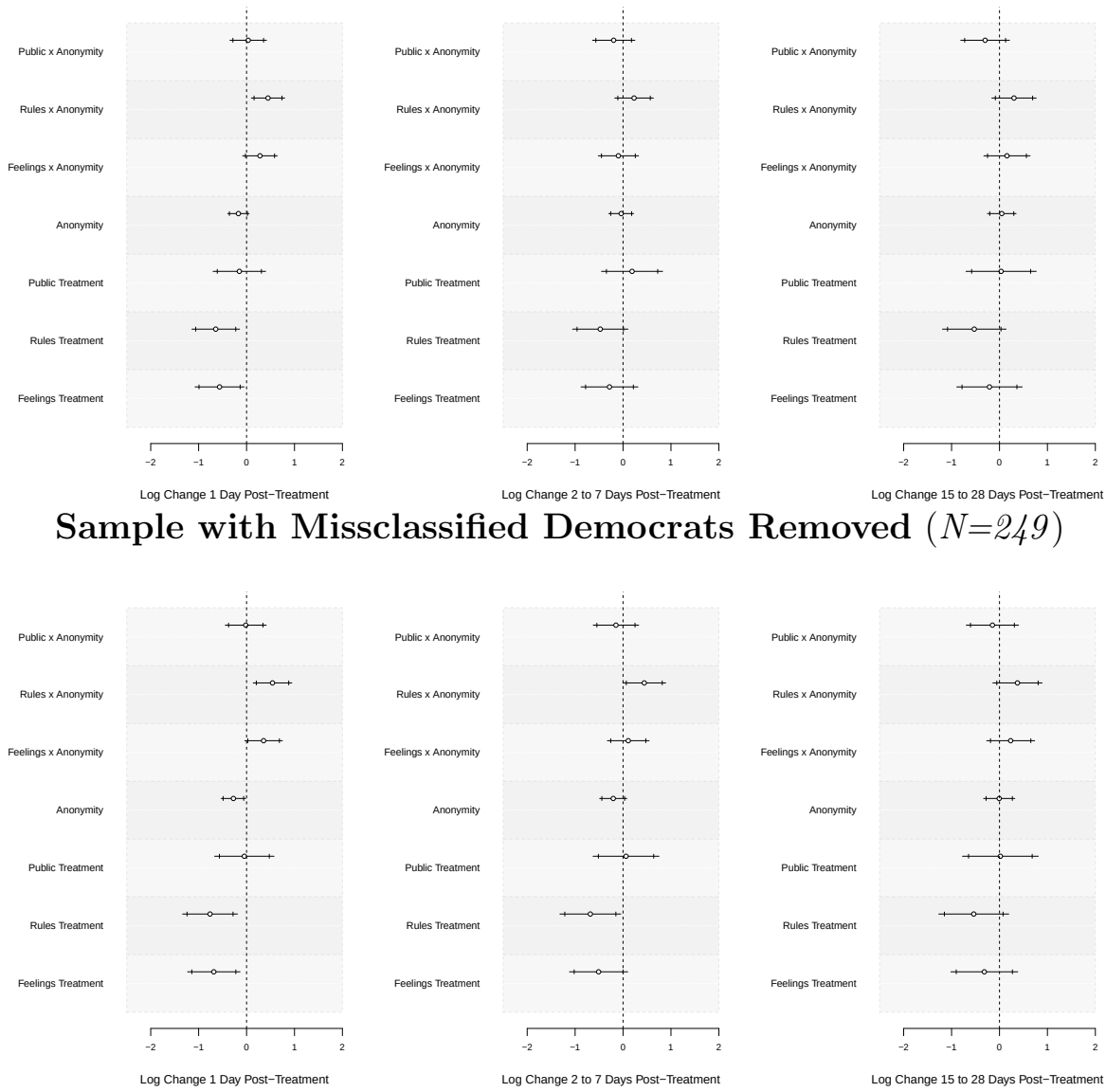
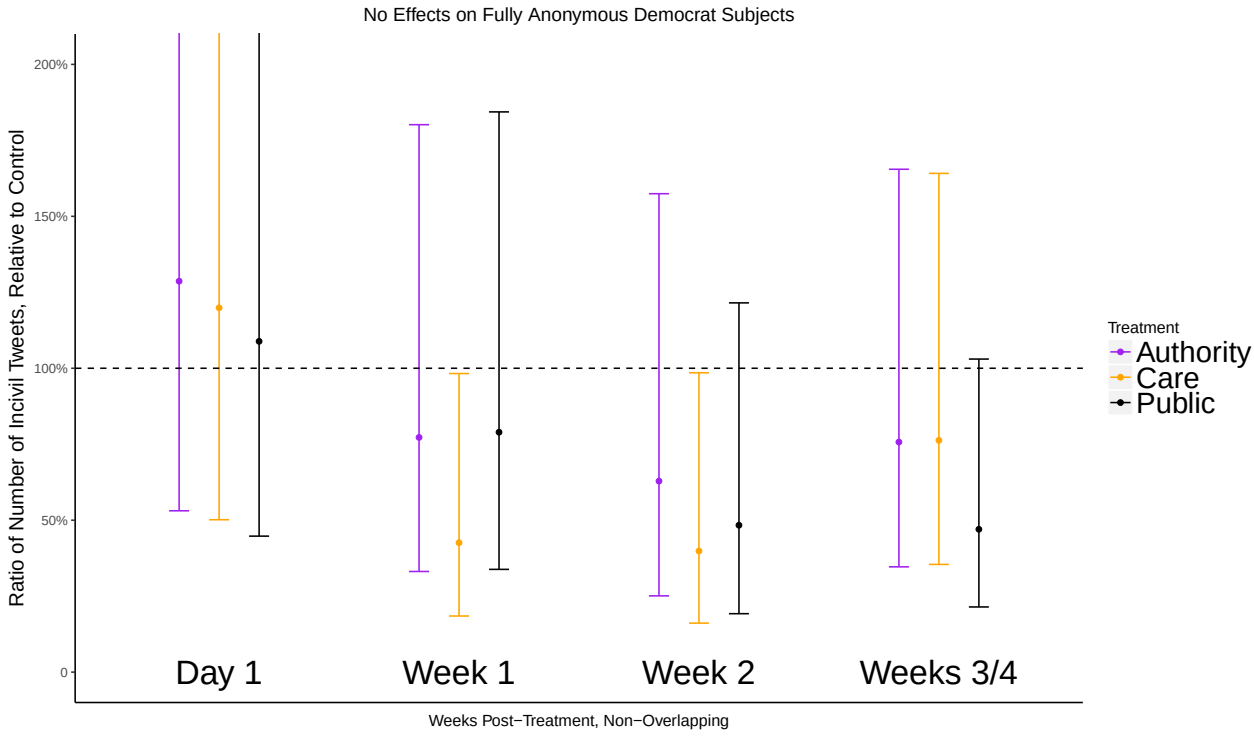


Figure 11: Each panel represents the results of a separate OLS regression in which the outcome variable is the log of the number of times a subject directed an incivil tweet at another user in the specified time period. The top three plots are calculated only on the Liberal sample, and the bottom three 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.

Figure 12: **Change in Incivility, Anonymous Democrat Sample** (*N=39*)



C Heterogeneous effects on partisan subsamples

Figure 13: **Change in Incivility, Semi-Anonymous Democrat Sample** (*N=48*)

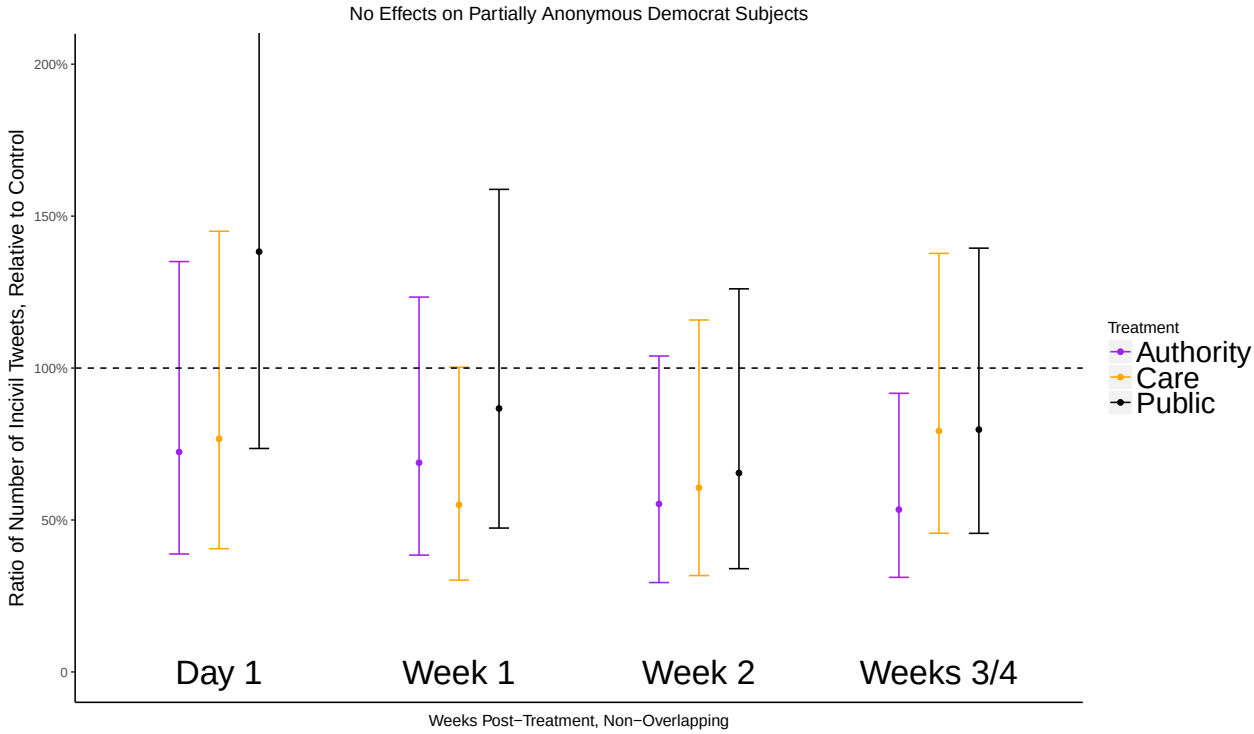


Figure 14: **Change in Incivility, Non-Anonymous Democrat Sample** (*N=60*)

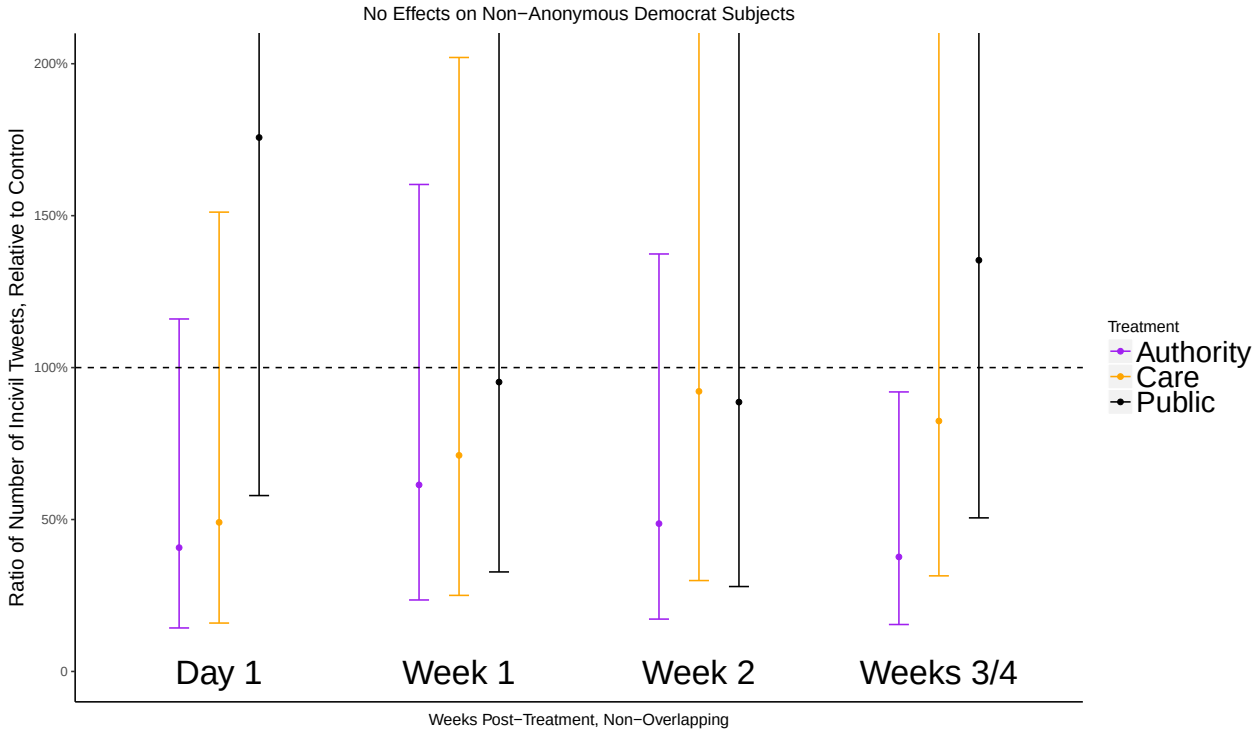


Figure 15: **Change in Incivility, Anonymous Republican Sample** (*N=44*)

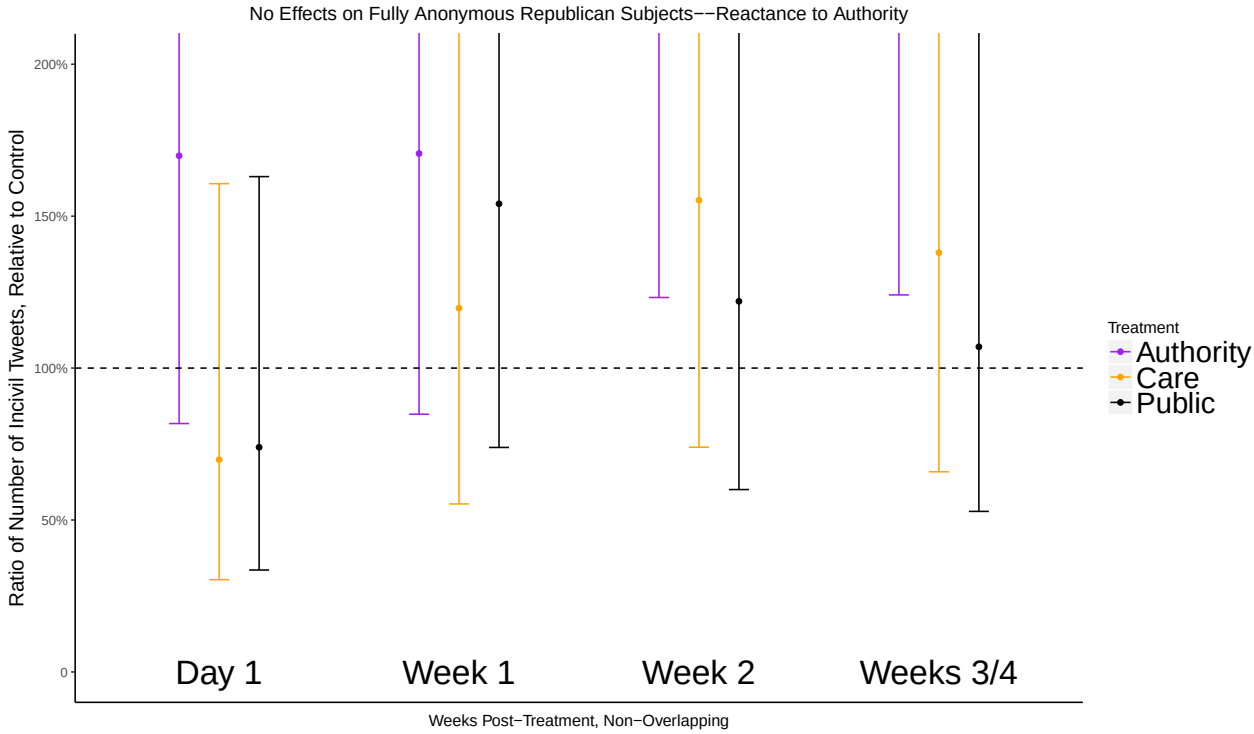


Figure 16: **Change in Incivility, Semi-Anonymous Republican Sample** (*N=46*)

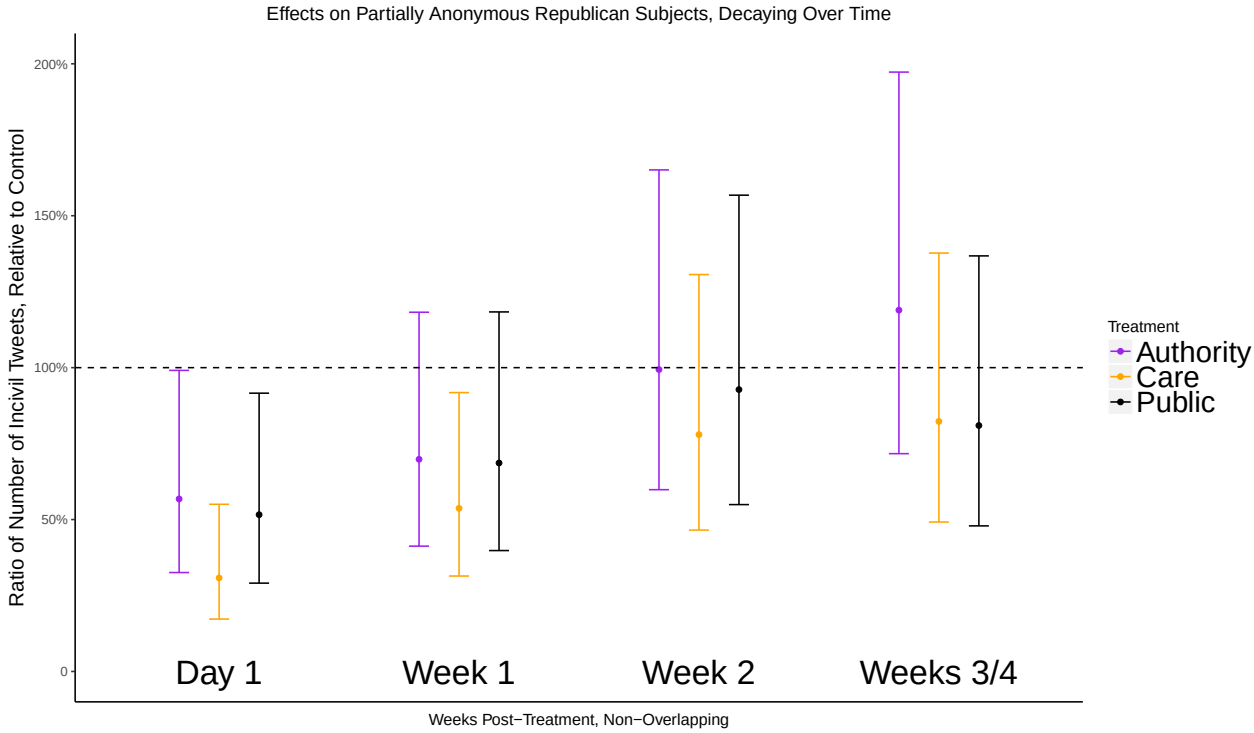
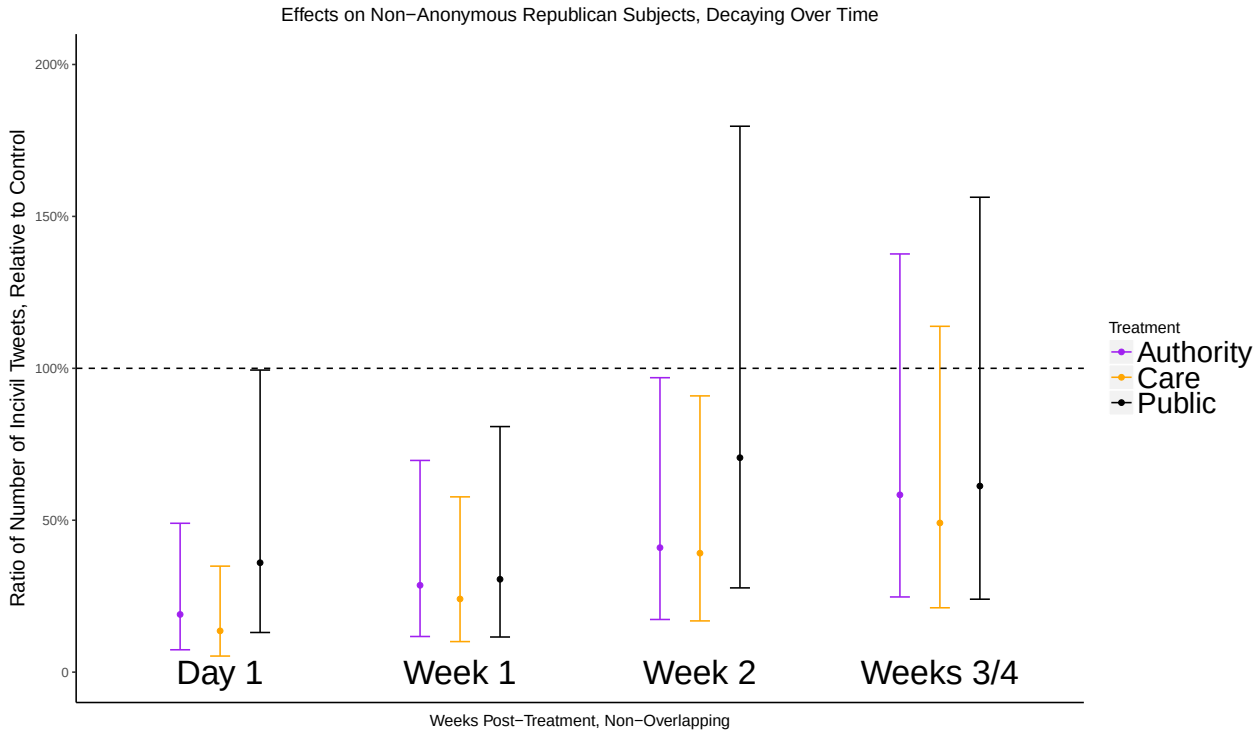


Figure 17: **Change in Incivility, Non-Anonymous Republican Sample** (*N=73*)



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.

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.".

Coleman, Gabriella. 2014. *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. Verso Books.

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.

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 A, and Neil Malhotra. 2017. "Political homophily in social relationships: Evidence from online dating behavior." *The Journal of Politics* 79 (1): 269–283.

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.

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. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." *Political Behavior* 39 (3): 629–649.
URL: *https://doi.org/10.1007/s11109-016-9373-5*

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.

Phillips, Whitney, and R Milner. 2017. "The ambivalent internet: Mischief, oddity, and antagonism online." *Hoboken, NJ: Wiley* .

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.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 1391–1399.