Supplemental Information: Table of Contents

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

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

Table 2: 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%

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

Finally, there were two accounts that were just weird; they had not tweeted thou- 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 Empirical distribution of aggression scores in subject tweets

As shown in Figure 9, the distribution of aggression scores (as coded by the algorithm developed by Wulczyn, Thain, and Dixon (2017)) is bimodal: there is a large cluster of "non-aggressive" tweets near 0, and a smaller cluster of "definitely aggressive" tweets near 1. The vertical line represents the 70th percentile of this empirical distribution, the cutoff I use in the body of the paper for transforming these scores into a binary measure. The higher cutoff of the 90th percentile would entail including only the far- right cluster of tweets. The main results are replicated using this higher threshold in Appendix D.

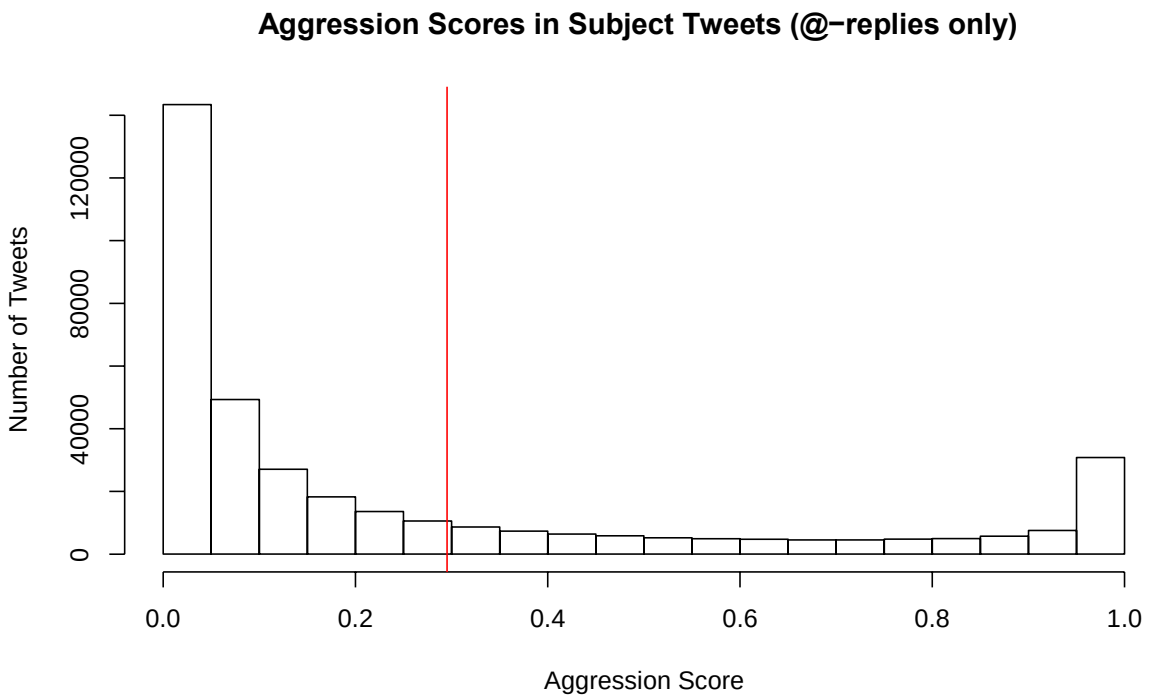


Figure 9: Empirical distribution of aggression scores. The vertical line represents the 70th percentile, the cutoff I use in the body of the paper.

C Validation of Wikipedia measure on the current dataset

Figure 10 plots the accuracy of the scores derived from the Wulczyn, Thain, and Dixon (2017) model in predicting the labels of tweets coded by crowdworkers. The x-axis plots the threshold used to turn the continuous scores output by the model into binary labels. There is a slight peak (accuracy = .82) at the black vertical line, which depicts the 75th percentile, but the accuracy is fairly constant across a wide range of cutoffs.

The validation tweets consist of 1,000 tweets which were randomly sampled from among all subject tweets and uploaded to Mechanical Turk. Each tweet was coded by two of Amazon's "Expert Coders," a restrictive label that they only award to consis- tently attentive crowdworkers. The precise instructions given to the workers were as follows:

Please read each tweet and tell us if it is civil or incivil.

We say that "civil" tweets are those that demonstrate respect for the person being tweeted at.

If a tweet has very little information (if it just contains a link, for example), code it as "civil."

Overall levels of intercoder reliability were low by the standards of objective clas- sification tasks (Krippendorf's alpha = .37). The task at hand, however, is inherently subjective, and our results are in line relevant published work: Wulczyn, Thain, and Dixon (2017), using a somewhat more rigorous coder vetting process, report "a Krip- pendorf's alpha score of 0.45. This result is in-line with results achieved in other crowdsourced studies of toxic behavior in online communities. (p3)"

For the accuracy results displayed in Figure 10, I restricted the initial 1,000 tweets to the 70% on which the coders agreed on the label. These labels are unbalanced in the sample (74% were labeled civil), so the 82% accuracy represents a significant improvement on a naive classification scheme.

As the confusion matrices below indicate, maximizing accuracy entails a tradeoff with balancing classification discrepancies. At the 70th percentile threshold, the percentage of validation tweets labeled as uncivil by the human coders but civil by the algorithm is 11.8%, compared to 6.5% for disagreements in the opposite direction.

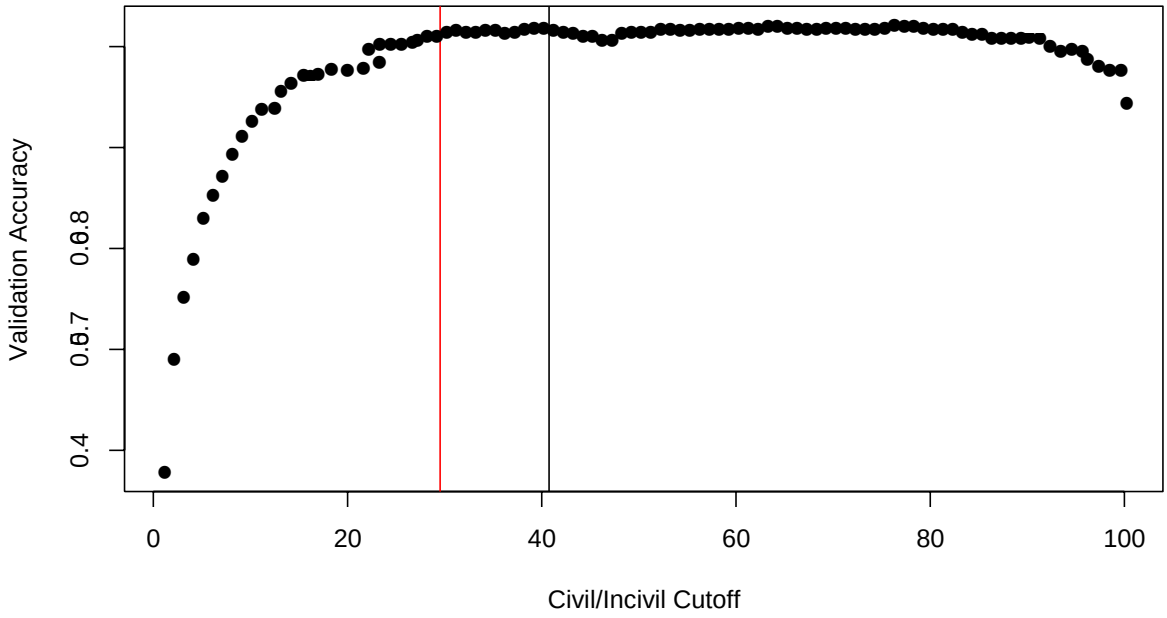


Figure 10: Accuracy of the Wikipedia model applied to tweets labeled by Mechanical Turk workers, scored on the tweets on which coders agreed on whether the tweet should be labeled civil or incivil. The red vertical line represents the 70th percentile, the cutoff I use in the body of the paper.

Confusion matrix of human and algorithmic labels on validation tweets; 75th percentile

	Mturk: Civil	Mturk: Incivil
Algorithm: Civil	67.7%	11.8%
Algorithm: Incivil	6.5%	13.9%

Confusion matrix of human and algorithmic labels on validation tweets; 70th percentile

	Mturk: Civil	Mturk: Incivil
Algorithm: Civil	65.3%	9.8%
Algorithm: Incivil	9.0%	15.9%

To balance these discrepancies, the second confusion matrix reports the results when the incivility threshold is lowered to the 70th percentile. This balancing comes at the cost of lowering the overall accuracy from 81.8 to 81.2. On balance, I believe that this slight decrease in accuracy is less important than using a measure that concords with the human coding as much as possible, so I use the 70th percentile threshold in the body of the text.

D Main Results Using Higher Aggression Threshold

The results in the body of the paper use the 70th percentile of the empirical distribution of aggression scores (as coded by the algorithm developed by Wulczyn, Thain, and Dixon (2017)) to code subject tweets as civil or incivil. There is a distinct cluster of "definitely aggressive" tweets near the top of this distribution, and the results in Figure 11 plot the model results when only this cluster is coded as incivil—that is, using the 90th percentile as the threshold. In both plots, the effects in the 1 Day time period become more pronounced to 0, while the effects in the 1 Week time period become closer to 0 and not statistically significant.

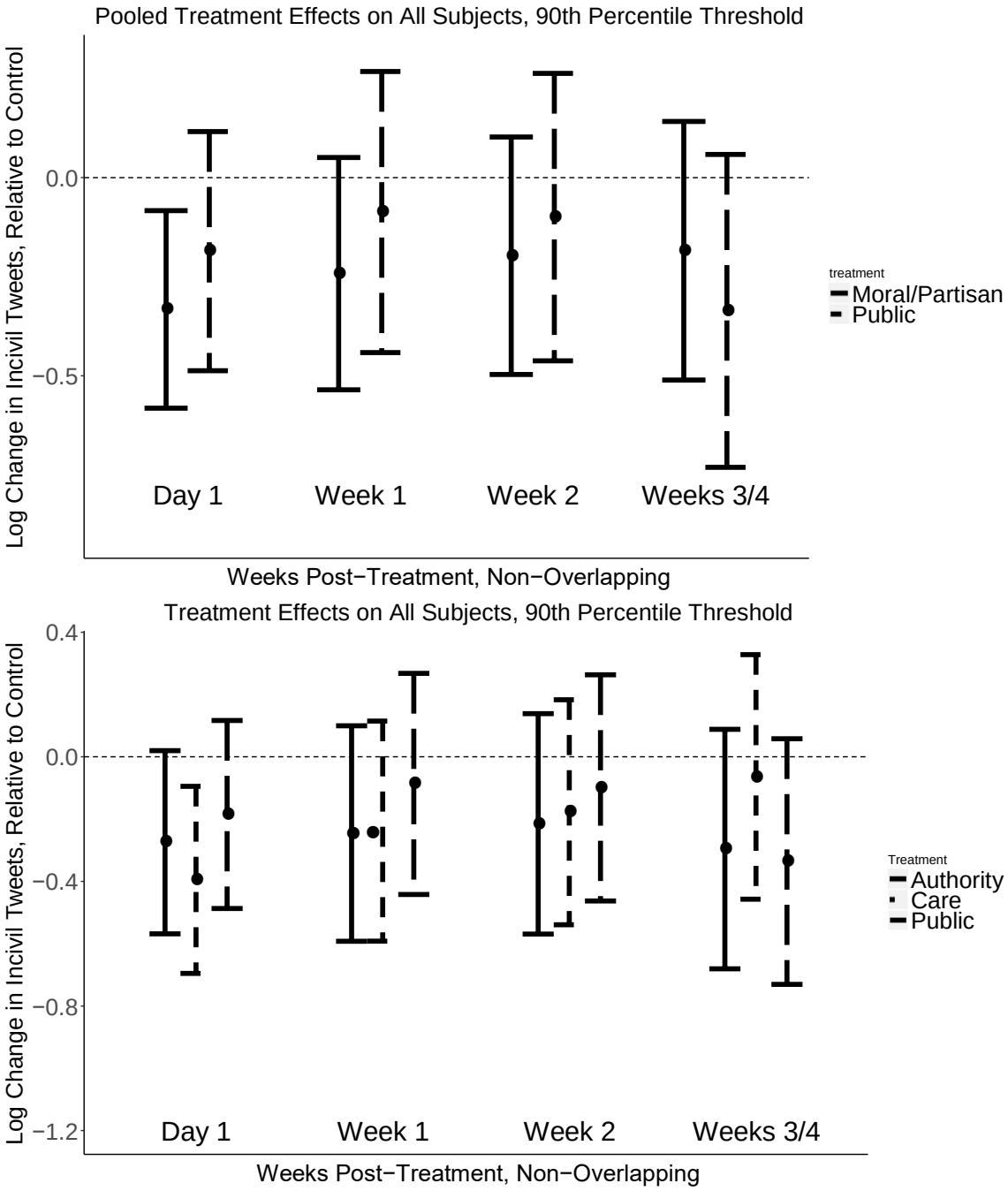


Figure 11: Main results replicated using the higher threshold of aggression scores for coding tweets as incivil.

E Negative Binomial Specification of Main Results

The dependent variable of interest in this analysis is the number of times a subject sent an uncivil tweet to another user. This is a "count variable"–it can only take non-negative integer values–and thus violates a fundamental assumption of OLS regression. To address this issue, generalized linear models with different assumptions are often used. Poisson regression, in which the dependent variable is assumed to have a Poisson distribution, is a common technique, but this carries the further assumption that the variance and expected value of the dependent variable are equal. In cases in which the variance is significantly higher than the expected value–like it is here–the negative binomial model relaxes this assumption (Hilbe, 2011).

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

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

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, 2011). For example, the effect of the Feelings treatment on subjects with Anonymity Score 1 (the middle category) is:

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

The results of these negative binomial models can be seen in Figure 12 and Figure 13.

Change in Incivility, Full Sample, Negative Binomial Specification (*N=310*)

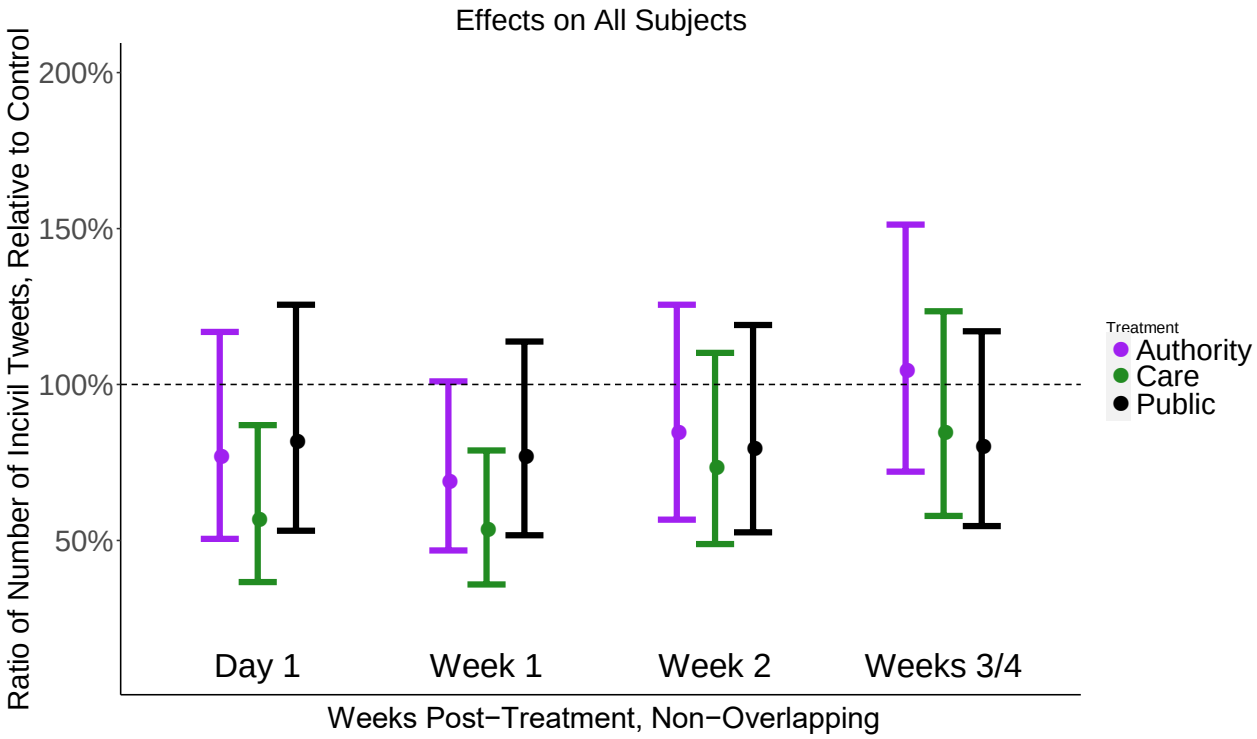
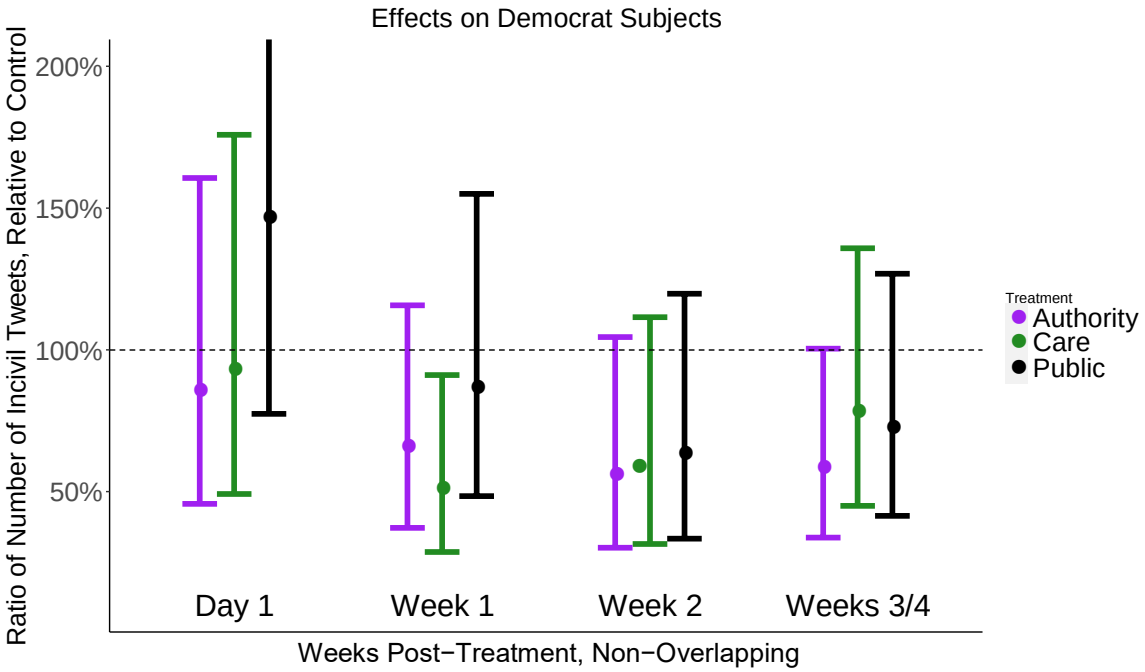
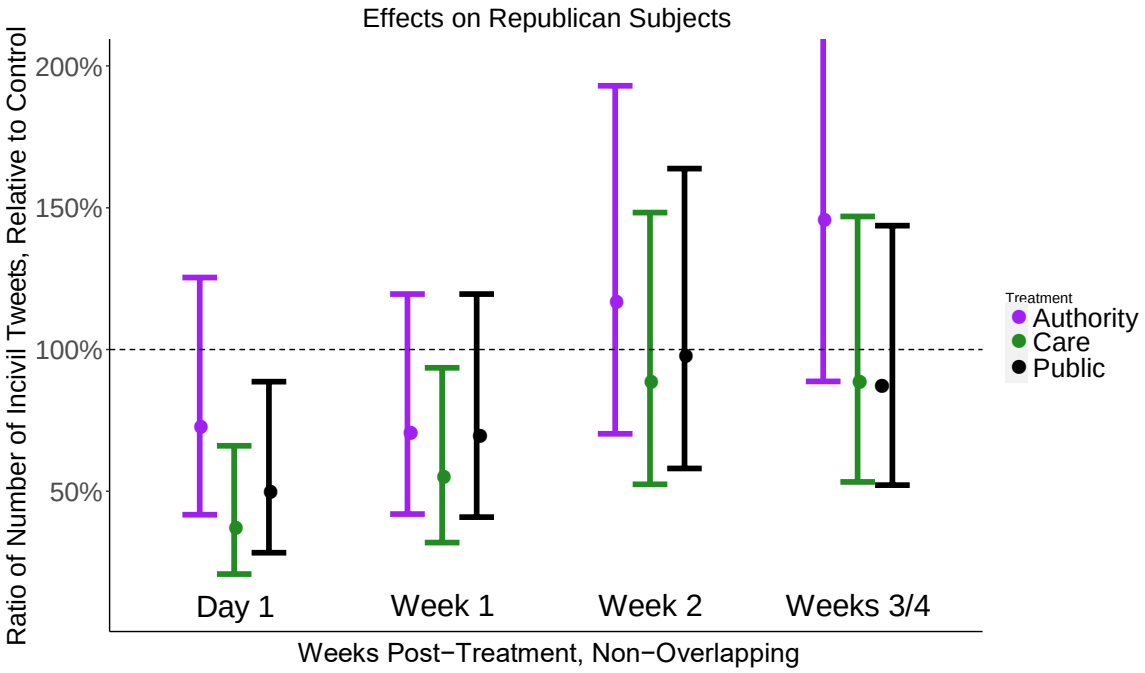


Figure 12: The Incidence Ratio calculated from the negative binomial model. For example, the Incidence Ratio associated with the Care treatment on Day 1 on the left of the plot means that these subjects sent 50% as many directed uncivil tweets as the subjects in the control group. 95% confidence intervals.

Figure 13: **Effects on Democrats, Negative Binomial Specification** (*N=147*)



Effects on Republicans, Negative Binomial Specification (*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 uncivil tweets as the subjects in the control group. 95% confidence intervals.

F Results Divided by Subject Loquacity

The subject population is highly varied in their pre-treatment level of tweeting activity. Although not one of the tests I specified in my pre-analysis plan, the potential policy implications of heterogeneous effects based on subject loquacity merit investigation of this possibility.

Figure 14 replicates the main results in the paper by the pre-treatment tweeting rate of the subjects. The top panel displays the results for subjects above the median (82 uncivil pre-treatment tweets), and the bottom panel for subjects blow this threshold.

There is a clear distinction: treatment effects on the more active subjects are close to zero, while the effects (of the moral treatments) on the less active subjects are significant in the 1 week time period.

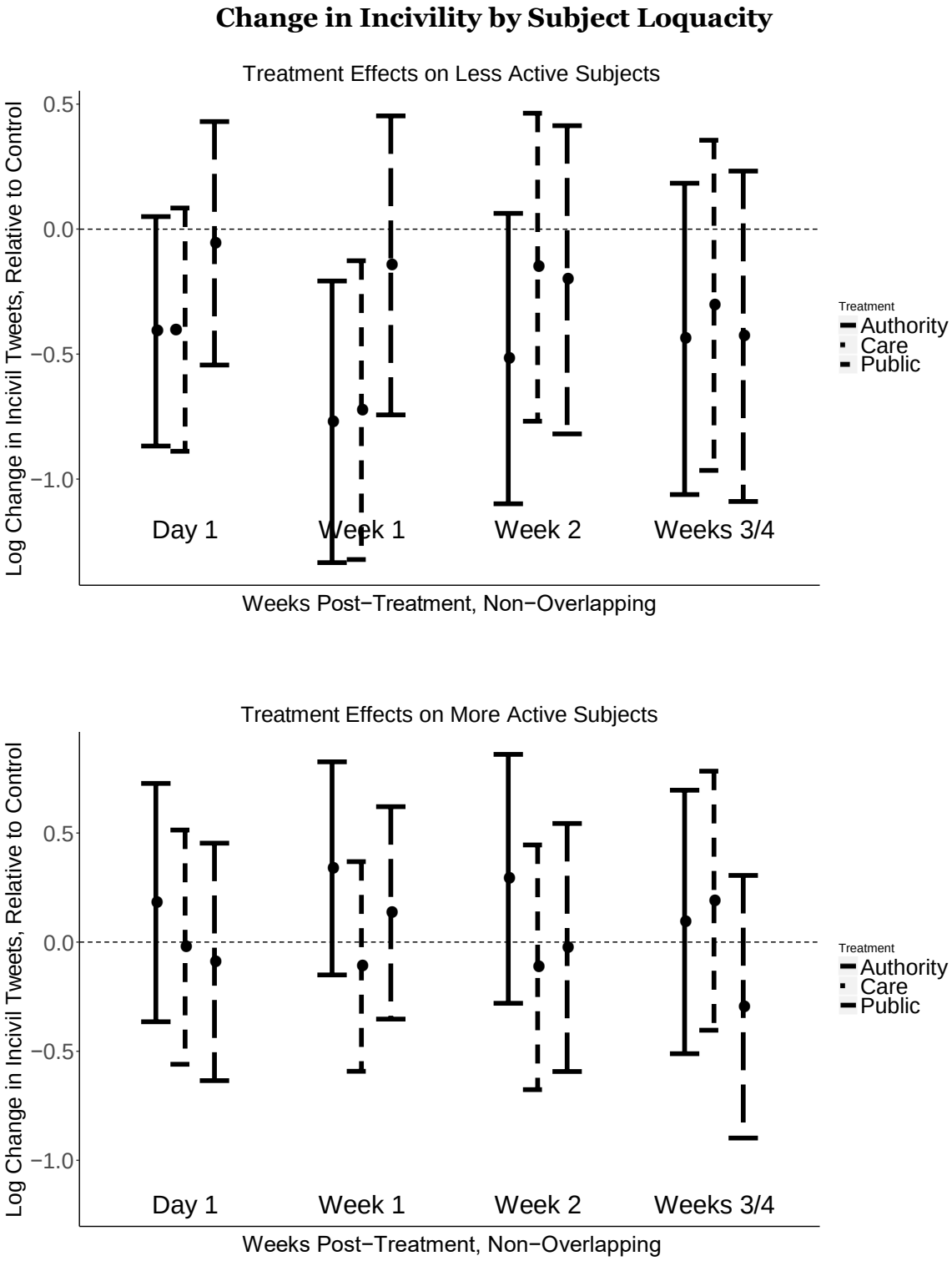


Figure 14: Treatment effects divided by subject pre-treatment tweeting rate. The top panel displays results for less active subjects (below the median), while the bottom panel displays results for more active subjects (above the median). Lines represent 95% confidence intervals.

G Treatment Effects on Sending Civil Tweets

The results in the body of the paper display treatment effects on the rate of sending *incivil* tweets. It is worth exploring whether the treatment had an analogous effect on sending *civil* tweets.

I re-ran the analysis using the number of *civil* tweets as the dependent variable (those with aggression scores below the 70th percentile threshold), and found no significant treatment effects. The point estimates are in the same direction as the effects on uncivil tweets, with effect sizes ranging from 50% to 80% as large. Figure 15 displays these results. In Panel A, examining pooled treatment effects, these effect sizes are for the civil tweets, .08 (1 Day) and .16 (1 Week); for the uncivil tweets in the body of the text, these effect sizes are, .15 and .2.

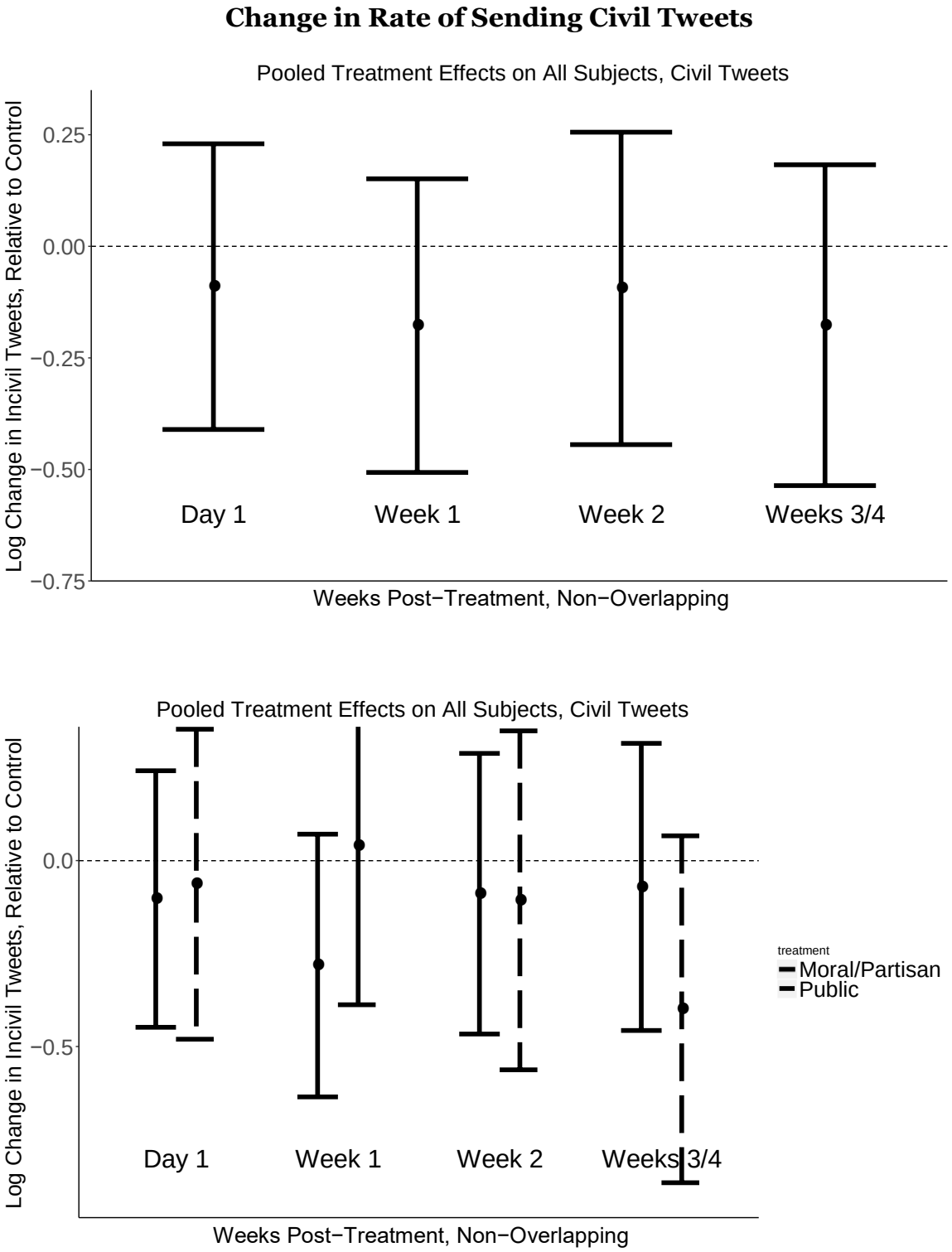
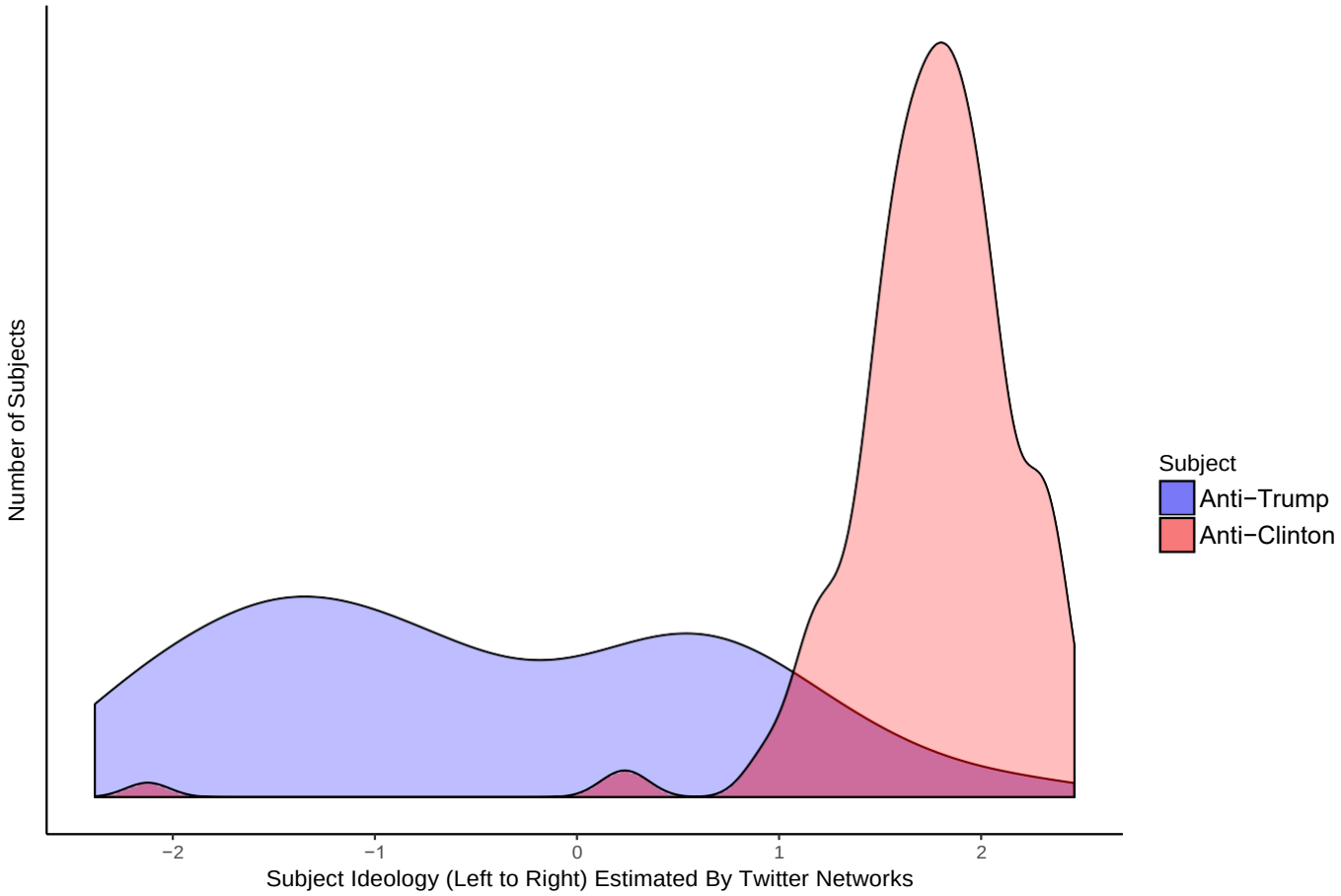


Figure 15: Treatment effects on the rate of subjects sending civil tweets (tweets scored as below the threshold for incivility used the body of the paper). The top panel displays results pooled across all treatment conditions, while the bottom panel displays results where the two moral treatments are pooled. Lines represent 95% confidence intervals.

Figure 16: Estimated Ideology of Subjects Labeled “Republican” or “Democrat”



H Ideological Heterogeneity

I implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points. As Figure 16 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 ideology scores above 1. 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 16, there appears to be two distinct clusters of Anti-Trump subjects. In addition to the expected group of Democrats, there is also a significant contingent of moderate

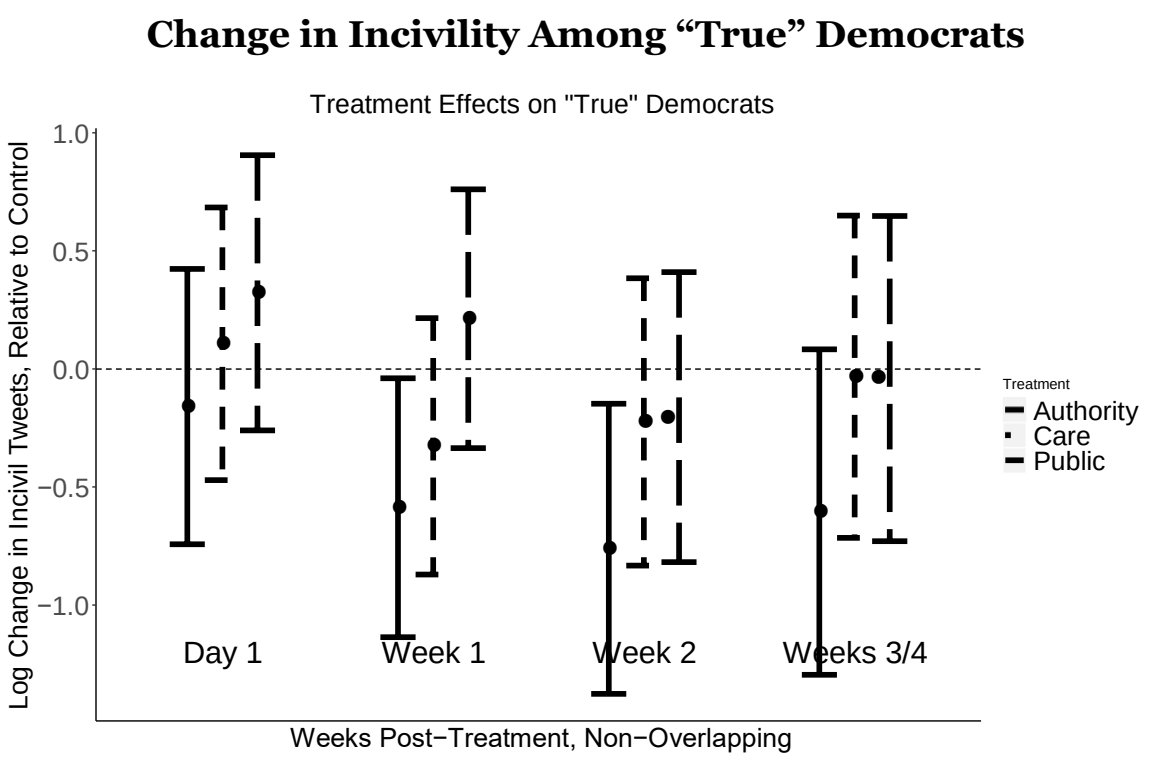


Figure 17: Treatment effects on Democrat subjects, restricted to subjects whose ideologies were estimated to be left of the anti-Clinton cluster on the right (*N=122*).

Anti-Trump Republicans that I classified as Democrats. Because the Care and Authority 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 7 to only those with estimated ideology scores to the left of the major cluster of anti-Clinton subjects in Figure 16, I find some support for this *ex post* explanation. The point estimates for the Authority treatment effect becomes more negative in the Week 1 and Week 2 time periods, seen in Figure 17. Because the sample size is down to 86, the Care treatment is still not significant, but the largest change is on the Authority effects, which are now significantly negative in the Week 1 and Week 2 time periods.

References

Barberá, Pablo. 2014. "streamR: Access to Twitter Streaming API via R." *R package version 0.2* 1.

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

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.

Chen, Adrian. 2015. "The Agency." *New York Times Magazine* June 2, 2015.
URL: *https://www.nytimes.com/2015/06/07/magazine/the-agency.html*

Cresci, Stefano, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Mau- rizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Com- mittee pp. 963–972.

Duggan, M, and A Smith. 2016. "The political environment on social media." *Pew Research Center* .
URL: *http://www.pewinternet.org/2016/10/25/political-content-on-social-media/*

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.

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 religion and politics.” *Pantheon, New York* .

Hilbe, Joseph M. 2011. *Negative binomial regression*. Cambridge University Press.

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

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

Muddiman, Ashley. 2017. “Personal and public levels of political incivility.” *International Journal of Communication* 11: 21.

Munger, Kevin. 2017. “Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment.” *Political Behavior* 39 (3): 629–649.

Munger, Kevin. 2020. “Replication Data for: Don’t @ Me: Experimentally Reducing Partisan Incivility on Twitter.”.
URL: *https://doi.org/10.7910/DVN/OUYTUP*

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

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

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

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.