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

Kevin Munger
New York University

1. Introduction

Partisan incivility is common in online discourse, and it poses a problem for democratic deliberation. I conduct an experiment which tests different interventions that promote a more civil discourse, combating trends that lead to lower levels of political learning and higher levels of partisan affect polarization.

Figure 1: *Non-Elite Incivility*

2. Theory: Partisan Affect, Online Incivility and Moral Suasion

- Partisans dislike each other more and more (“affect polarization”)
- Competition for attention incentivizes outrageous and uncivil behavior (Hindman, 2008)
- Partisan incivility decreases political learning and reinforces affect polarization (Mutz, 2015)
- Zero barriers to entry and anonymity mean that civility norms on Twitter are very poor
- Moral suasion requires sympathy and shared moral foundations (Haidt, 2012)
- Motivation for “field” experiments to promote civility
  - ~~Experiments in the lab~~: Experiment in the “field”:
  - ~~Convenience samples~~ Sample of real, consistently uncivil users
  - ~~Short time frame~~ Continuous and unbounded time frame
  - ~~In the lab~~ In the same context as the uncivil political discussion
- Goal: **use Twitter bots to send messages that sanction uncivil discourse and cause subjects to use less of it**

3. Sending Messages From Partisan Bots Designed to Look Like Real Accounts

4. Intervening in Uncivil Discussion with Partisan Bots

“Sean” is a subject in my experiment: his tweet to “NSA Agent Jim” is an example of political incivility between non-elites with different partisan leanings. Because “Sean” was uncivil to someone who was uncivil to Trump, I label “Sean” as a Republican and sent him a message from a Republican bot.

5. Measuring Partisan Incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
- Model is a multi-layer perceptron using character-level n-grams; extremely black box, but more accurate than a single coder
- Pre-registered, not related to the existing data
- Classifies each tweet on a 0 to 1 scale; code each tweet as uncivil if it is on the right tail of the distribution of aggression scores (75th percentile; robust to using the 70th or 80th percentile)
- Outcome measure is the number of uncivil tweets each subject sends post-treatment

6. Hypotheses

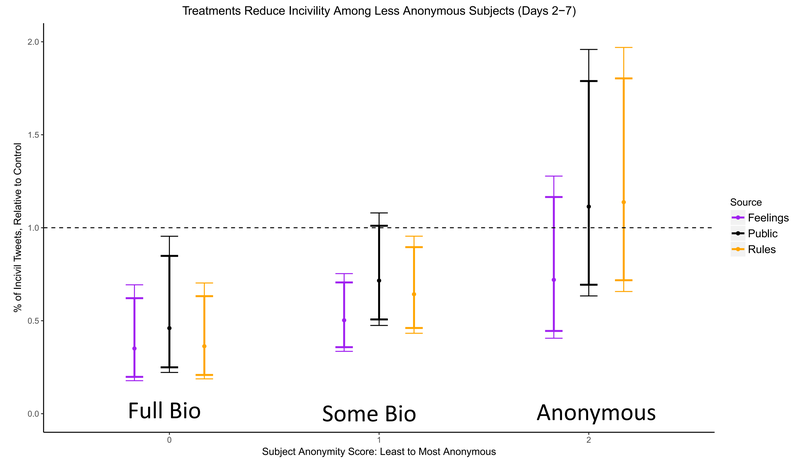
Pre-registered through EGAP (number 20150520AA)

Hypothesis 1. Subjects who receive one of the sanctioning messages will send fewer uncivil tweets.

- In the Feelings condition, this effect will be larger for liberals than for conservatives.
- In the Rules condition, this effect will be larger for conservatives than for liberals.
- The effect of the Public condition will be symmetric but smaller than in the other conditions.

Hypothesis 2. The reduction in incivility caused by the sanctioning messages will be larger for more anonymous subjects.

7. Results



8. Anti-Trump Subjects are Ideologically Diverse; Anti-Hillary Subjects are Not