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

Kevin Munger

NYU

August 29, 2017
Prepared for Twitter 2017
Partisan incivility is bad for democracy and especially common online.
Project Outline

- Partisan incivility is bad for democracy and especially common online
- Test interventions aimed at discouraging partisan incivility
Partisan incivility is bad for democracy and especially common online

Test interventions aimed at discouraging partisan incivility

Use bots to send randomly-assigned messages with varied moral appeals

- Feelings treatment
- Rules treatment
- Public treatment
Partisan incivility is bad for democracy and especially common online

Test interventions aimed at discouraging partisan incivility

Use bots to send randomly-assigned messages with varied moral appeals

- Feelings treatment
- Rules treatment
- Public treatment

Track changes in the rate of incivility relative to a control group
Finding political incivility

@realDonaldTrump you already lost

@realDonaldTrump fuck you dirt bag

All around nice guy!
Who Cares?

- Twitter...is important for US politics
Who Cares?

- Twitter...is important for US politics
- High levels of “affect polarization” in the offline electorate (Iyengar & Westwood, 2015)
Who Cares?

- Twitter...is important for US politics
- High levels of “affect polarization” in the offline electorate (Iyengar & Westwood, 2015)
- Exposure to incivility leads to greater affect polarization and less information acquisition (Mutz 2015)

Mutz (2015): “uncivil discourse is communication that violates the norms of politeness of a given culture...Following the rules of civility/politeness is...a means of demonstrating mutual respect.”

Incivil discourse may make deliberative democracy impossible

Participants sincerely weigh the merits of arguments, regardless of who makes them (Fishkin, 2009)
Who Cares?

- Twitter...is important for US politics
- High levels of “affect polarization” in the offline electorate (Iyengar & Westwood, 2015)
- Exposure to incivility leads to greater affect polarization and less information acquisition (Mutz 2015)
- Mutz (2015): “uncivil discourse is communication that violates the norms of politeness of a given culture...Following the rules of civility/politeness is...a means of demonstrating mutual respect.”

Incivil discourse may make deliberative democracy impossible. Participants sincerely weigh the merits of arguments, regardless of who makes them (Fishkin, 2009).
Who Cares?

- Twitter...is important for US politics
- High levels of “affect polarization” in the offline electorate (Iyengar & Westwood, 2015)
- Exposure to incivility leads to greater affect polarization and less information acquisition (Mutz 2015)
- Mutz (2015): “uncivil discourse is communication that violates the norms of politeness of a given culture...Following the rules of civility/politeness is...a means of demonstrating mutual respect.”
- Incivil discourse may make deliberative democracy impossible
Who Cares?

- Twitter...is important for US politics
- High levels of “affect polarization” in the offline electorate (Iyengar & Westwood, 2015)
- Exposure to incivility leads to greater affect polarization and less information acquisition (Mutz 2015)
- Mutz (2015): “uncivil discourse is communication that violates the norms of politeness of a given culture...Following the rules of civility/politeness is...a means of demonstrating mutual respect.”
- Incivil discourse may make deliberative democracy *impossible*
  - Participants sincerely weigh the merits of arguments, regardless of who makes them (Fishkin, 2009)
Political incivility online

- Computer-mediated communication lacks biological feedback
Political incivility online

- Computer-mediated communication lacks biological feedback
- Competition for attention online: “It may be easy to speak in cyberspace, but it remains difficult to be heard” (Hindman, 2008)
Political incivility online

- Computer-mediated communication lacks biological feedback
- Competition for attention online: “It may be easy to speak in cyberspace, but it remains difficult to be heard” (Hindman, 2008)
- Small number of committed bad actors (“trolls”), on eg 4chan (Phillips, 2015)
Political incivility online

- Computer-mediated communication lacks biological feedback
- Competition for attention online: “It may be easy to speak in cyberspace, but it remains difficult to be heard” (Hindman, 2008)
- Small number of committed bad actors (“trolls”), on eg 4chan (Phillips, 2015)
- Incivility can cause people to disengage from politics on social media—even politicians (Theocharis et al, 2016)
Political incivility online

- Computer-mediated communication lacks biological feedback
- Competition for attention online: “It may be easy to speak in cyberspace, but it remains difficult to be heard” (Hindman, 2008)
- Small number of committed bad actors (“trolls”), on eg 4chan (Phillips, 2015)
- Incivility can cause people to disengage from politics on social media—even politicians (Theocharis et al, 2016)
- Seeing uncivil comments can cause a wider group of people to act uncivilly (Cheng et al, 2017)
Manipulating political discourse

- Experiments in the lab
  - Convenience samples
  - Short time frame
  - In the lab
Manipulating political discourse  My Approach

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
Finding political incivility

- Needs to be fast, and accurate
Finding political incivility

- Needs to be fast, and accurate
- Can afford to compromise on recall
Finding political incivility

- Needs to be fast, and accurate
- Can afford to compromise on recall
- Only interested in:
  - real (non-elite) users
Finding political incivility

- Needs to be fast, and accurate
- Can afford to compromise on recall
- Only interested in:
  - real (non-elite) users
  - arguing with out-partisans
Finding political incivility

- Needs to be fast, and accurate
- Can afford to compromise on recall
- Only interested in:
  - real (non-elite) users
  - arguing with out-partisans
  - about politics
Finding political incivility

Parker @ [black] 21h
@realDonaldTrump you already lost

Ty [black] @ [black] 9:32 AM - 17 Oct 2016
@ [black] @realDonaldTrump fuck you dirt bag
StreamR finds a tweet with "@realDonaldTrump" or "@HillaryClinton"

Is the tweet an "@"-reply to someone besides Trump or Clinton?

EXCLUDE

Calculate aggression score; is tweet in top 10% most aggressive?

EXCLUDE

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

EXCLUDE

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

EXCLUDE

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

EXCLUDE

Assign to a treatment condition subject to balance constraints
A Visual Overview

Donald J. Trump @realDonaldTrump - Oct 11
In Texas now, leaving soon for BIG rally in Florida!

NSA agent jim @[redacted] - Oct 11
@realDonaldTrump HOPE U DIE

Sean [redacted] @[redacted]
@[redacted] @realDonaldTrump typical libtard go fuck yourself
11:25 AM - 12 Oct 2016

Matthew [redacted] @Matthew [redacted] Oct 12
@Matthew [redacted] You shouldn't use language like that. Republicans need to remember that our opponents are real people, with real feelings.
Measuring incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
Measuring incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
- Trained on over 100,000 human-coded comments
Measuring incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
- Trained on over 100,000 human-coded comments
- Model is a multi-layer perceptron using character-level n-grams; extremely black box, but more accurate than a single coder
Measuring incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
- Trained on over 100,000 human-coded comments
- 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
Measuring incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
- Trained on over 100,000 human-coded comments
- 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
- Restricted subject tweets (367k) to those directed “@” another user

Classified a tweet as uncivil if score $> 75$th percentile (robust to 70th or 80th).
Measuring incivility

- Use a machine learning model to evaluate “aggressiveness” in Wikipedia comments (Wulczyn, Thain & Dixon, 2017)
- Trained on over 100,000 human-coded comments
- 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
- Restricted subject tweets (367k) to those directed “@” another user
- Classified a tweet as a uncivil if score > 75th percentile (robust to 70th or 80th)
Treatment Variations

- Moral intuitionist model (Haidt, 2001): moral emotion is antecedent to moral reasoning

- Moral appeals should target intuitions rather than (epiphenominal) logic
- Different rhetoric to appeal to different moral frameworks
  - Authority moral foundation: "You shouldn't use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility."
  - Care moral foundation: "You shouldn't use language like that. [Democrats/Republicans] need to remember that our opponents are real people, with real feelings."
  - Non-moral message: "Remember that everything you post here is public. Everyone can see that you tweeted this."
Treatment Variations

- Moral intuitionist model (Haidt, 2001): moral emotion is antecedent to moral reasoning
- Moral appeals should target intuitions rather than (epiphenomenal) logic
Treatment Variations

- Moral intuitionist model (Haidt, 2001): moral emotion is antecedent to moral reasoning
- Moral appeals should target intuitions rather than (epiphenomenal) logic
- Different rhetoric to appeal to different moral frameworks
Treatment Variations

- Moral intuitionist model (Haidt, 2001): moral emotion is antecedent to moral reasoning
- Moral appeals should target intuitions rather than (epiphenominal) logic
- Different rhetoric to appeal to different moral frameworks
  - Authority moral foundation: “You shouldn’t use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility.”
Treatment Variations

- Moral intuitionist model (Haidt, 2001): moral emotion is antecedent to moral reasoning
- Moral appeals should target intuitions rather than (epiphenomenal) logic
- Different rhetoric to appeal to different moral frameworks
  - Authority moral foundation: “You shouldn’t use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility.”
  - Care moral foundation: “You shouldn’t use language like that. [Democrats/Republicans] need to remember that our opponents are real people, with real feelings.”
Treatment Variations

- Moral intuitionist model (Haidt, 2001): moral emotion is antecedent to moral reasoning
- Moral appeals should target intuitions rather than (epiphenomenal) logic
- Different rhetoric to appeal to different moral frameworks
  - Authority moral foundation: “You shouldn’t use language like that. [Democrats/Republicans] need to behave according to the proper rules of political civility.”
  - Care moral foundation: “You shouldn’t use language like that. [Democrats/Republicans] need to remember that our opponents are real people, with real feelings.”
  - Non-moral message: “Remember that everything you post here is public. Everyone can see that you tweeted this.”
Treatment uptake
Treatment uptake
Hypotheses

Hypothesis pre-registered through EGAP.

**Hypothesis**

*The effect of the Care condition will be larger for liberals than for conservatives. There will be an effect of the Authority condition for conservatives, but not for liberals. There will be an effect of the Public condition, but it will be smaller than the other effects.*
Hypotheses

Hypothesis pre-registered through EGAP.

**Hypothesis**

*The effect of the Care condition will be larger for liberals than for conservatives. There will be an effect of the Authority condition for conservatives, but not for liberals. There will be an effect of the Public condition, but it will be smaller than the other effects.*

**Hypothesis**

*Treatment effects will be smaller for more anonymous subjects.*
Results–responses to interventions

Response Rates by Treatment (N=224)
Results—responses to interventions

Response Rates by Anonymity

- Full Info (N=62)
- Some Info (N=78)
- Anonymous (N=84)
Results—responses to interventions

Percentage of Conciliatory Response (N=72)

<table>
<thead>
<tr>
<th></th>
<th>LEFTIST</th>
<th>RIGHTIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feelings</td>
<td>(0/14)</td>
<td>(3/8)</td>
</tr>
<tr>
<td>Rules</td>
<td>(1/11)</td>
<td>(5/17)</td>
</tr>
<tr>
<td>Public</td>
<td>(3/8)</td>
<td>(1/11)</td>
</tr>
</tbody>
</table>
Negative Binomial Specification

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

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

\[
V_{feel \times Anon_1} = V(\hat{\beta}_2) + Anon^2 V(\hat{\beta}_6) + 2Anon \times Cov(\hat{\beta}_2, \hat{\beta}_6)
\]
Change in Incivility, Full Sample \((N=310)\)

Effects on All Subjects, Declining Over Time

- Day 1
- Week 1
- Week 2
- Weeks 3/4

Ratio of Number of Incivil Tweets, Relative to Control

- Authority
- Care
- Public

Weeks Post-Treatment, Non-Overlapping
Change in Incivility, Anonymous Sample ($N=133$)

No Effects on Fully Anonymous Subjects

Weeks Post-Treatment, Non-Overlapping
Change in Incivility, Semi-Anonymous Sample, No Interaction Effects ($N=94$)

Effects on Partially Anonymous Subjects, Decaying Over Time

Ratio of Number of Incivil Tweets, Relative to Control

- **Authority**
- **Care**
- **Public**

Weeks Post-Treatment, Non-Overlapping

- Day 1
- Week 1
- Week 2
- Weeks 3/4
Change in Incivility, Non-Anonymous Sample \((N=83)\)

Effects on Non–Anonymous Subjects, Decaying Over Time

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Weeks 3/4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio of Number of Incivil Tweets, Relative to Control</td>
<td>0%</td>
<td>50%</td>
<td>100%</td>
<td>150%</td>
</tr>
</tbody>
</table>
Change in Incivility, Republican Sample, No Interaction Effects ($N=163$)

Effects on All Republican Subjects

Weeks Post-Treatment, Non-Overlapping

Ratio of Number of Incivil Tweets, Relative to Control

- Authority
- Care
- Public

Day 1 | Week 1 | Week 2 | Weeks 3/4

0 | 50% | 100% | 150% | 200%
Change in Incivility, Democrat Sample, No Interaction Effects ($N=147$)
Anti-Trump Subjects were Ideologically Diverse

Subject Ideology (Left to Right) Estimated By Twitter Networks

Number of Subjects

Subject

Anti-Trump

Anti-Clinton

Subject Ideology

-2 -1 0 1 2

Number of Subjects

-2 -1 0 1 2

Subject

Anti-Trump

Anti-Clinton

Diagram labels:
- Subject
  - Anti-Trump
  - Anti-Clinton

Diagram legend:
- Anti-Trump
- Anti-Clinton
“Real” Democrat Sample, No Interaction Effects ($N=86$)

Effects on “Real” Democrat Subjects

Weeks Post-Treatment, Non-Overlapping

Ratio of Number of Incivil Tweets, Relative to Control
Encouraging Online Civility

- Large and persistent treatment effects (on most of the sample) from a single intervention
Encouraging Online Civility

- Large and persistent treatment effects (on most of the sample) from a single intervention
- Small, persistent groups promoting incivility online
Encouraging Online Civility

- Large and persistent treatment effects (on most of the sample) from a single intervention
- Small, persistent groups promoting incivility online
  - Trolls
  - Ideologues
- My hope: most people would prefer civility
Thanks for your comments, and for listening!

km2713@nyu.edu
@kmmunger (please be civil)
## Attrition rates

**Table:** Attrition Rates and Causes

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