Electoral Studies 62 (2019) 102072

Contents lists available at ScienceDirect

Electoral Studies

journal homepage: http://www.elsevier.com/locate/electstud

Tweeting for peace: Experimental evidence from the 2016 Colombian Plebiscite

Jorge Gallego [a], Juan D. Martínez [b], Kevin Munger [c], Mateo Vásquez-Cortés [d,*]

[a] Universidad del Rosario, Colombia
[b] Duke University, USA
[c] Penn State University, USA
[d] ITAM, USA

A B S T R A C T

The decades-long Colombian civil war nearly came to an official end with the 2016 Peace Plebiscite, which was ultimately defeated in a narrow vote. This conflict has deeply divided Colombian civil society, and non-political public figures have played a crucial role in structuring debate on the topic. To understand the mechanisms underlying the influence of members of civil society on political discussion, we performed a randomized experiment on Colombian Twitter users shortly before this election. Sampling from a pool of subjects who had been frequently tweeting about the Plebiscite, we tweeted messages that encouraged subjects to consider different aspects of the decision. We varied the identity (a general, a scientist, and a priest) of the accounts we used and the content of the messages we sent. We found little evidence that any of our interventions were successful in persuading subjects to change their attitudes. However, we show that our pro-Peace messages encouraged liberal Colombians to engage in significantly more public deliberation on the subject.

1. Introduction

Recent elections throughout the world have raised concerns on whether political actors, both domestic and international, are using bots to shape political conversation and alter electoral behavior (Murthy et al., 2016).[1] Allegations of this kind have been made in various places, such as the U.S. (Bessi and Ferrara, 2016), the U.K. (Howard and Kollanyi, 2016), Germany (Brachten et al., 2017), Catalonia (Alandete, 2017), France (Ferrara, 2017), among many others. Latin American countries are not exempt from these practices, as the evidence suggests that in recent elections in Venezuela, Mexico, and Brazil, bots have been actively used to manipulate public opinion. Initial optimism surrounded the consolidation of the internet in general, and of social media in particular, as many argued that these platforms would open new spaces for democratic deliberation and would give voice to many that were previously excluded by traditional media. However, the consolidation of some pervasive methods and tactics, such as fake news, misinformation, bots, trolls, among others, whose clear intention is to shape public opinion in favor of certain actors, has undermined the expectations surrounding this type of platforms (Tucker et al., 2017).

Recent episodes, such as the Cambridge Analytica scandal, have called into question the preparedness of governments and technological companies in preventing and responding to these strategies. Not surprisingly, recent research has attempted to understand the scale of activity and the impact of bots on political behavior (Stukal et al., 2018). However, most of the existing studies focus on the problem of identifying bots (Chu et al., 2012; Ferrara et al., 2016; Stukal et al., 2018) and on understanding the type of behavior followed by these accounts (Murthy et al., 2016). In contrast, very few studies center on the consequences that bots may have on political preferences and voting behavior. Understanding if bots can alter citizens' preferences and behavior is crucial in order to determine what type of actions governments should take to prevent these practices.

Motivated by this debate, in this paper we study if social media accounts represent-ing non-political public figures can affect citizens' attitudes and behaviors in a context of high polarization and in the midst of an important election. Our goal is to under-stand if bots representing public figures may persuade subjects to endorse a particular political position. Moreover, we want to understand if these potential effects vary along different relevant dimensions, such as the identity of the account, the content of the messages sent by the bot, and the ideology of the recipients of these messages. What type of speaker would have a greater impact for different segments of the ideological spectrum? What combination of identities and messages are more effective in changing

* Corresponding author.
E-mail address: mv1093@nyu.edu (M. Vásquez-Cortés).

[1] Bots are automated accounts that post content based on algorithms (Tucker et al., 2017).

https://doi.org/10.1016/j.electstud.2019.102072
Received 7 January 2019; Received in revised form 7 June 2019; Accepted 1 September 2019
Available online 17 October 2019

deliberative decisions, such as participating more in the debate?

We address these questions in the context of the 2016 Colombian Peace Plebiscite.[2] We test if messages related to the peace agreement reached by the Colombian government and the guerrilla group FARC, from like-minded social media accounts representing important public figures, cause positive reactions and increased engagement from citizens. We test this causally by conducting an experimental study using Twitter "bots" that we control (Munger, 2017a,b). This approach allows us to perform the experiment on the sample of interest—Colombian Twitter users who frequently posted comments about the plebiscite—in a naturalistic setting.

After more than 50 years of war, the Colombian government and the FARC reached a peace accord. Citizens had the choice to approve reject this deal through a plebiscite that took place on October 2nd, 2016. Society was very polarized at the time of the election. As in recent elections in other contexts, social media proved to be one of the most important platforms to express opinions on both sides of the debate. Facebook and Twitter were used heavily to exchange opinions about the peace accord the weeks before the election. Even though the case of a peace plebiscite in a country affected by several decades of war seems to be idiosyncratic, we believe that this example shares some common and important traits with other past and future elections, in which the society is deeply divided around an issue and public figures can shape and encourage deliberation through social media.

The experiment was conducted on Colombian Twitter users shortly before the election. Sampling from a pool of subjects who had been frequently tweeting about the Plebiscite, we tweeted messages that encouraged subjects to consider different aspects of the decision. In doing so, we test whether bots are able to cause subjects to engage in more discussion on this topic—and whether they can change subjects' expressed sentiment towards the Plebiscite—to better understand the mechanisms underlying the potential influence of these strategies on political expression and online behavior.

Existing evidence suggests that ideology shapes people's response to information and affects how they make political decisions (Campbell, 1960), but in this process several factors are crucial. Moral convictions, for example, shape political attitudes (Graham et al., 2009; Lakoff, 2002; Morgan et al., 2010). Consequently, in order to convince someone to adopt a certain position or to take a particular political action, it is necessary to appeal to the counterparts' moral convictions— sometimes called "moral reframing" (Feinberg and Miller, 2015; Volkel and Feinberg, 2017). If a liberal wants to convince a conservative about a certain issue, she must elaborate an argument based on the moral values in which conservatives believe. Our bots and their messages were designed to test some of these theories.

On the other hand, several scholars have argued that deliberation is essential for any democracy, as it prepares citizens for further political action. Accordingly, deliberation can increase levels of political knowledge, civic engagement, and tolerance (Coleman and Blumer, 2009; Eveland, 2004; Graham, 2015; Johnston et al., 2001; Kim et al., 1999). However, it is still unclear whether social media increase the levels of deliberation and political talk in a society. Evidence from Facebook (Robertson et al., 2010), Twitter (Yardi and Boyd, 2010), Weblogs (Papacharissi, 2009), Youtube (Halpern and Gibbs, 2013), and Wikipedia (Black et al., 2011) are decidedly mixed. To some degree, one of the goals of using bots in the midst of an election is to increase public talk on a particular topic or candidate.

In this paper, we investigate whether bot influence in the form of a single message from an account that appears to be a potentially influential figure in the Colombian society (a priest, a scientist or a general) can change the way people think and talk about this important political decision. Hence, our study represents a *proof of concept*, in the sense that we do not model the exact way in which bots are utilized—usually through botnets and a massive number of messages—but instead, we simplify the intervention in order to understand if a single message sent by a bot produces any visible effect on citizens' preferences, attitudes, and online behavior.

This paper represents a contribution to the burgeoning literature that explores the effects of social media on political outcomes, and in particular, on how it might under-mine democracy (Tucker et al., 2017). Several studies have explored how social media may exacerbate certain pathologies like disinformation, fake news, echo chambers, political polarization, incivility, among others. Within this context, we provide a rigorous test of the effects that bots may have on political attitudes and online behavior. We believe that our findings represent an important contribution, for several reasons.

First, most studies on bots have focused on methods for identifying them and describing their behavior, while few studies try to understand whether they exert any influence on citizens' expressions and behavior. Among the few studies to ask these kinds of questions is Bail et al. (2018), who conduct an experiment in the US in which a large group of Democrats and Republicans are paid to follow bots retweeting information from accounts with clear opposing political views. The study shows that efforts to expose citizens to opposing views may backfire, as some republicans become more conservative and some Democrats more liberal when confronted with opposing views.

Also, Murthy et al. (2016) analyze how bots may influence conversational networks on Twitter, in the context of the UK general election in 2015. For this purpose, the authors recruited student volunteer participants to create new Twitter accounts in charge of commenting high-profile broadcast events. Bots where linked to a random partition of these accounts, while the rest where not affected by bots, in order to determine if the former would become more influential than the latter. The authors find no significant differences between the two groups, concluding that accounts linked to bots are not more influential. Our study differs from Bail et al. (2018) and Murthy et al. (2016) in that we do not limit ourselves to expose subjects to bots and opposing information, but we also test if the identity of the bot and the content of the messages have any effects on the expressive behaviors of subjects.

Our second contribution lies on the fact that most studies on social media and political expression are restricted to developed countries. However, understanding if bots may affect preferences and behavior in developing countries is important, because in these countries voters tend to be, on average, less educated and have lower levels of information. These are precisely the contexts where manipulation is expected to be most effective. However, our results show that, even in this case, shifting the attitudes expressed on social media is difficult; at most, bots increase the conversation among those that were aligned with the fake account from the beginning.

Finally, we contribute to the debate on the ethics of online experiments. We acknowledge that using fake accounts to study public opinion in the context of an important election is not without risks. However, we argue that we have taken the appropriate steps to minimize these risks. Novel areas of human behavior require novel research designs, and we believe that scholars should not shy away from studying controversial topics. Research ethics does not require an absolute minimizing of risk, but rather a careful consideration of the balance of risks and benefits. Given the importance of the global discussion surrounding the electoral influence of bots, we believe that our study well exceeds that standard.[3]

We find little evidence this intervention caused anyone to modify their expressed attitudes—very few people switched the sentiment of their tweets towards the peace process as a result, which is somewhat

[2] A plebiscite is the direct vote of all members of an electorate on an important question pertaining an official policy.

[3] Further, we are not violating Twitter's terms of service nor any Colombian law, and we have not come anywhere close to altering the results of any election—if such alteration were possible given the moderate scale of our intervention, global democracy would be in a fragile place indeed.

unsurprising given our highly motivated sample and the low intensity of our treatment. However, we do find changes in public discussion on this topic. We find that liberals (who advocated for the peace agreement) were motivated to send more messages in favor of the process after receiving a favorable message. Conservatives did not send more of these positive messages, but neither did they send more negative messages. Hence, a variety of bots representing cultural figures were able to spur increased participation in the online discussion of this important political event, albeit in an unexpected way, providing evidence of the "confirmation bias" theorized to describe political interaction in social media. We conclude that it is unlikely for bots to *alter* citizens' online expression, but that is possible for them to *amplify* existing patterns of expression.

2. The peace plebiscite, Twitter and public figures

The referendum under consideration consisted of a single question that voters had to approve or reject: "Do you support the final agreement to end the conflict and build a stable and lasting peace?" Two sides campaigned during the weeks preceding the referendum. The 'Yes' campaign was supported by the political left, center-left and center, led by President Juan Manuel Santos. The most prominent campaigner for the 'No' vote was the Centro Democratico, a right wing party led by current senator and former president Alvaro Uribe.

We must be careful whenever we extrapolate our findings to the behavior of the country as a whole. We acknowledge that the group of Twitter users talking about the peace process is not a representative sample of the Colombian population. However, we also believe that this population of Twitter users is an interesting population by itself, insofar it becomes more decisive and influential every day. Moreover, this election provided an ideal opportunity to analyze the effects that bots may have on sentiments and opinions about the Colombian peace process on Twitter. The referendum was conducted without explicit party labels on the ballot and concerned the single most important issue in Colombian politics.

We focus on a sample of Twitter users for several reasons. First, Twitter is growing as an important source of communication between citizens and the political elite in Colombia. Second, Twitter is a particularly important platform if we want to study the expansion of deliberation and public talk. Early optimistic views of social media argued that they allowed for direct communication without geographical obstacles. However, they quickly became a focus of incivility, in which offensive speeches abound (Buckels et al., 2014). During the Plebiscite, Twitter received special attention that highlighted the intensity and incivility of the debate.[4] The conversation was both constant and plagued with insults and disrespect. Therefore, the combination of its growing relevance and the heated tone of the conversation that takes place in the platform, makes a study that focuses on Twitter users particularly relevant.

3. Experimental design

We conducted a field experiment on Twitter during the 2016 Colombian Plebiscite.[5] We first collected all tweets related to the peace process from March through September of 2016. We identified accounts that were been active on this topic two months prior to the plebiscite.[6]

We then selected a random sample of 4500 of these accounts.[7] Using block randomization, with two blocks differentiating between supporters and opponents, we constructed seven groups (six treatment groups and a control group).

The actual experimental manipulation was to send public messages to subjects. All of the messages were in favor of the peace process, for reasons we discuss below. We varied the treatment on two dimensions: the identity of the sender and the ideological framing of the message. To manipulate identity, we created "bots" that had public profiles identifying them as one of three figures: a general, a priest, or a scientist. Fig. 1 shows the accounts of the liberal scientist and the conservative general.[8]

We sent two types of messages: a conservative message that emphasized typical conservative values such as patriotism, authority, and sanctity; and a liberal message that emphasized liberal values such as harm, fairness, and reciprocity. We rotated through the bots and tweeted the messages:

> Conservative: **"@[subject] The peace agreement is a victory of our compatriots and the will of God. Prosperity awaits for our homeland**"
>
> Liberal: **"@[subject] This war has taken 260,000 lives and 5 million missing. The poor suffer more. We can stop this**"

4. Ethics of online field experiments

Note that while these two messages emphasize different values, they both argue in favor of the peace process. Although we could have varied this dimension and included messages that argued against the peace process, ultimately, we decided not to. Including this variation would have meant that our treatments would have varied in three ways: type of bot, content of the message, and political position of the message. We would have needed a larger sample size to support this $3 \times 2 \times 2$ design, or sacrifice one of the other two dimensions. Given the hypotheses that we wanted to test, we opted to sacrifice the political position dimension and keep it constant throughout our treatments. There is also an ethical consideration to this decision—all else equal, it is better to minimize the amount of deception involved in an experiment like this, and we were ourselves in favor of the peace process. The messages we sent were factual, and we would happily have sent them from personal accounts; the only deception was in the identities of the bots.[9] Political scientists have frequently conducted research on the effectiveness of persuasion in the context of door-to-door canvassing and advertising; Kalla and Broockman (2018) identify 49 such experiments, and note that "most experiments in the literature have been conducted with Democratic or liberal-leaning organizations." This is because most American political scientists are liberal-leaning or Democrats; the norm in the discipline is for researchers to work for causes which they personally support.

The ethics of field experimentation is currently a point of debate in political science, especially in the context of large-scale election-related studies (Desposato, 2015, 2016). We believe that the current study is defensible on the primary grounds being discussed: the risk that we would influence the outcome of an election was zero, and the detailed measurement strategy of our research design meant that our sample was in the thousands rather than the hundreds of thousands of larger-scale

[4] See, for instance, this news coverage: https://colombiacheck.com/datos/especiales/la-guerra-se-traslada-a-twitter.html.

[5] We registered the Pre-Analysis Plan of this project at the Evidence in Governance and Politics (EGAP) platform. See http://www.egap.org/registration/2136.

[6] Details of this process can be found in Appendix D.

[7] Fig. 9 in Appendix describes the selection process of the accounts that were ultimately used in the experiment.

[8] The accounts were created and manipulated to make them look as real as possible. We bought followers to each and programmed them to constantly tweet about other issues related to their identity but unrelated to the peace process or politics in general. In fact, no messages were received with allegations that the accounts were fake, spam, or bots. On the contrary, many subjects responded to the messages as if the accounts were real.

[9] The research described in this paper was approved by the IRB at NYU and Universidad de Rosario.

Fig. 1. Treatments—scientist (liberal message) and general (conservative message).

voting studies.

One important dimension on which our current protocol does not conform with the guidelines suggested in Desposato (2015) is that we do not debrief our subjects. Our concern at the time was that doing so would "poison the well" of the subject population, making debriefed subjects permanently skeptical of political Twitter messages from strangers. In the light of recent reports of the use of fake partisan Twitter accounts to sow discord in the 2016 US election (Rosenberg, 2018), our position on this has changed: this well should be poisoned. We encourage future researchers employing similar research designs to debrief subjects.

The revelation of the existence of these fake accounts (the extent of which was only made clear after our field research was completed) has increased the salience of discussions of fake accounts on social networks, and raises questions about the ethical implications of using fake accounts for any research activities. One crucial consideration here is illustrated by the different rules enforced by Twitter and Facebook relating to the identities of accounts on their networks.

Facebook requires all users to use their real name; parody or comedy accounts are not allowed. Conducting an analogous experiment to the one described in this manuscript on Facebook would entail a fundamental violation of Facebook's core aim of ensuring that all accounts are associated with exactly one real person, and we would argue that such an experiment would be more ethically fraught. Twitter, however, permits the creation of accounts that do not correspond to real people; there is a rule against impersonating real individuals, but our accounts do not violate the Twitter Rules (as of July 18, 2018).

One way in which our design may violate the Twitter Terms of Service is in the purchasing of fake followers for our bots to increase their legitimacy.[10] This deception is unavoidable; Munger (2017b) shows that messages from Twitter accounts with low followers have no effect (or can even backfire). Munger reports (potentially) violating the Terms of Service in just this way in this paper. There is also a precedent for the practice of purchasing Twitter followers to understand the dynamics of this practice in Computer Science; see, for example, Shah et al. (2017) and Aggarwal and Kumaraguru (2015).

[10] The Twitter Rules are inconclusive on this point. "Spam: You may not use Twitter's services for the purpose of spamming anyone … Some of the factors that we take into account when determining what conduct is considered to be spamming include: ….if you sell, purchase, or attempt to artificially inflate account interactions (such as followers, retweets, likes, etc.)." Twitter appears to be defending their ability to enforce rules against "spam" at will, rather than explicitly defining the behavior of buying followers as being a *prima facie* violation of their rules.

Furthermore, we argue that the discipline of political science should not give more weight to the wishes of a private company like Twitter than to, say, the government of China. King, Pan, and Roberts (2014)—an influential paper, published in *Science*—report creating 2 fake accounts on each of 100 Chinese discussion websites. These fake accounts wrote 2 posts each (either pro-government or anti-government) in pub-lic forums to see which would be censored by the Chinese government; that is, they intentionally broke Chinese law to test patterns of Chinese law enforcement. Addition-ally, these posts were public; although there were only 400 posts, the population who was exposed to these posts was potentially in the millions. In contrast, the current experiment was targeted, and only visible to the treated subjects (unless they chose to retweet the messages to their followers).

On both dimensions—the magnitude of the rule violation and the normative legit-imacy of the rulemaker (or, in the Chinese case, *law-maker*)—the research design of King et al. (2014) represents a bigger concern than that of the current manuscript.

However, one improvement to our design for future research would be to replace the fake "bot" accounts with real people (confederates). These confederates would agree to send messages they agree with, minimizing deception and maximizing account verisimilitude. This design could also take advantage of the fact that the confeder-ates are embedded in genuine online networks. Designs like this have recently been used successfully (though at relatively small sample size) on Facebook (Haenschen, 2016). Nonetheless, it is important to mention that the election of the identities of our accounts—a general, a priest, and a scientist—does not violate any law in Colombia. For the general, which could be the most sensitive case in terms of impersonation, we refrained from using any official logos, badges, or references to particular military units. In any case, this issue did not raise any concerns to Universidad del Rosario's ethics committee in Bogota.

In general, a universal ethical maxim that entails following every rule of every entity governing online content is untenable; there are some research activities which are morally and ethically justifiable, and some which are not. These dimensions do not map directly onto the wishes of online rulemakers, so researchers need to be careful but confident that we can govern ourselves. This is a rapidly evolving area, and there do not yet exist disciplinary guidelines delineating permissible research activities.

Although the specific syntheses of our research design are relatively novel, both the researcher-as-partisan when sending partisan messages during a campaign and the creation-of-fake-accounts-on-social-media aspects of the design have numerous prece-dents in the literature.

5. Data

We first need to know whether the subject replied directly to the bot's tweet. This behavior indicates whether or not the subject accepts the message and sender as a legitimate authority, and could explain the mechanism by which future behavior changes. Overall, 158 subjects (4% of those in a treatment group) sent a tweet directly in reply to our bots. We coded these as either positive or negative reactions and analyze them in Section E of Appendix.

The primary behavior targeted in this experiment is the frequency and sentiment of tweets about the peace process. To capture this behavior, we scraped each subject's Twitter history before and after the treatment and restricted the sample to the tweets that were about the peace process. We used a conservative approach to identifying these tweets: a dictionary of popular phrases and hashtags. Any tweet containing one of these key terms was coded as being about the peace process.[11]

To control for each subjects' pre-treatment behavior, we calculated their rate of tweeting about the peace process in the three months before the experiment. This measure was included as a covariate in all of the following analysis.

To test our hypothesis that the effect of our treatments would be moderated by the ideology of the subjects, we need to be able to say which of subjects were liberal or conservative. We implemented the method developed by Barberá (2015) to estimate subjects' ideological ideal points. This model looks at the accounts that each account follows and iteratively updates the closeness of each account in the network. The main intuition behind the method is that the probability of following someone on Twitter increases with the ideological closeness between the two accounts. This was possible for 2741 of the 3516 subjects who followed enough Colombian political elites. Figs. 2 and 3 report the ideological scaling of political elites and subjects in Colombia.

There was a strong connection between ideology in the traditional left-right political spectrum and support for the peace process. We validate this conventional wisdom by analyzing the ideology of the parties involved in the campaign according to their Twitter networks. Fig. 2 plots these estimates for each of member of Congress. The estimates in Fig. 2 pass the test of face validity of the relative positions of the parties in the referendum campaign: Centro Democratico is located at the right of the graph, which represents a more conservative ideology and consequently a negative view of the referendum. President Santos' main coalition—Partido de la U, Cambio Radical and Liberal—are

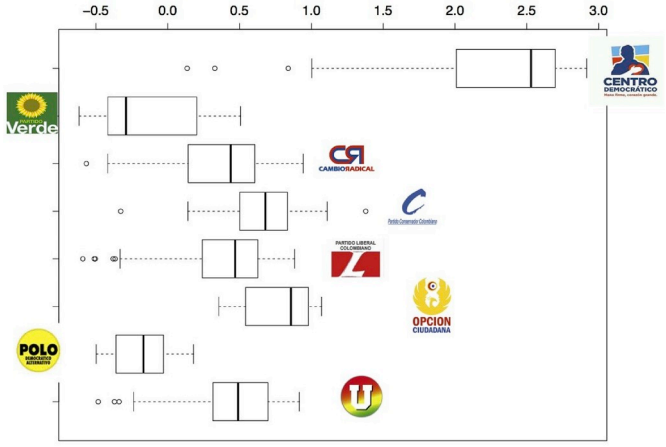

Fig. 2. Ideological Position of Political Parties in Colombia. Estimates of the ideological position of each Member of Congress in Colombia, sorted by their party. These estimates are derived from Barberá (2015)'s Bayesian Spatial Following Model.

[11] We selected the most popular hashtags related to the discussion of the peace pro-cess, as well as terms that tended to co-occur with those hashtags. There are undoubt-edly some tweets that we miss with this approach, but unless this classification covaries with our randomly assigned treatment, this should not be a problem for our analy-sis. These are the key terms we selected: "#AdiósALaGuerra","#PazCompleta","Acuerdo Gobierno FARC","Acuerdo FARC","Firma paz","paz Santos","proceso de paz","#PazenColombia","FARC Habana","paz Colombia","acuerdo Habana","#SíALaPaz","Uribe paz","plebiscito si","#ProcesoDePaz","#SiALaPaz","plebiscito paz","gobierno FARC","paz FARC","acuerdo paz","#FirmaDeLaPaz","#EnCartagenaDecimosNo","#FeliSídad","plebiscito no","#Plebiscito","diálogos de paz","#SantosElTal23NoExiste","Habana paz","negociaciones Ha-bana","conversaciones Habana","Gobierno paz FARC","Gobierno Habana FARC", and "#NoMarcho".

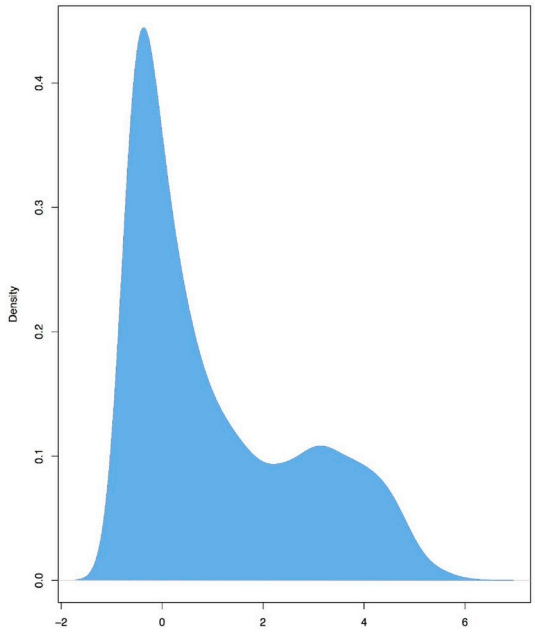

Fig. 3. Ideology Distribution of Subjects. Histogram of the ideological position of each of the subjects in our experiment. These estimates are derived from Barberá (2015)'s Bayesian Spatial Following Model.

located towards the center right of the scale. Polo Democratico and Partido Verde, which are the left parties in Colombia and supported the peace process, are located at the left of the graph.

After scaling members of the Congress, the method estimates the ideological scores of ordinary citizens. As seen in Fig. 3, the distribution of ideology of the Twitter users that were part of our experiment is slightly skewed. There is a tall, thin cluster of liberals and a broader cluster of conservatives, as well as a range of moderate individuals between them. Accordingly, we divide the political spectrum into three segments: liberal, conservatives, and moderates.

In addition to the raw number of tweets about the peace process, we were interested in their orientation: was the tweet in favor of "Si" or "No"? We call this the *sentiment* of the tweet. We began by hand-coding a balanced sample of 2000 tweets as in favor of "Si" (*pro*) or "No" (*con*). After pre-processing the text of the tweets, we then used a Naive Bayes (NB) classifier on these labeled tweets. NB is a commonly-used and computationally efficient machine learning technique; our model performed well, with a cross-validated out-of-sample prediction accuracy of 75%.[12] We then applied the trained NB model to the rest of the tweets, generating binary sentiment scores for the 70,000 subject tweets we identified as being about the peace process.

As a validity check of our application of these two machine learning classifiers, we plot the log (plus one) of the number of positive and negative tweets each subject sent about the peace process against their estimated ideology score, in Fig. 4. The top panel plots the log number of negative tweets; as expected, liberal subjects (with negative ideology scores) sent far fewer negative tweets about the peace process.

The bottom panel plots the log number of positive tweets, and the trend is reversed: liberal subjects sent more positive tweets about the peace process. In both cases, the trend is steepest (and the points densest) among subjects with ideology scores ranging from −1 to 1. This is because the ideology scores generated by the Barberá (2015) algorithm in this case exhibited a long right tail. Roughly half of the subjects were estimated to be on either side of the 0 midpoint, the correct proportion. However, the network of conservatives was much more segregated than that of the liberals, enabling the algorithm to give finer-grained estimates for extreme conservatives.

As a result, we cannot use these continuous ideology scores as covariates in the model: the marginal change in the subjects' ideology is not constant throughout the range of this variable. A change from −1 to 1 indicates a switch from a liberal to a moderate conservative subject, but a change from 4 to 5 indicates a switch from an ex-treme conservative to an even more extreme conservative. We thus create a categorical Ideology Score variable that takes the value 0 for subjects estimated below the 25th percentile ("Liberals"); the value 1 for subjects between the 25th and 75th percentile ("Moderates"); and the value 2 for subjects above the 75th percentile ("Conservatives").

6. Results

Our main analysis uses the subjects' of pre- and post-treatment tweets, categorized as discussed in the "Data" section above as pertaining to the peace process and being either positive or negative about the vote. These data are count data, so OLS would be inappropriate. A chi-squared test indicates that the counts are overdispersed, so following Munger (2017a), we used negative binomial regression.[13]

To interpret the relevant treatment effects implied by the coefficients estimated by these models, 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 Ideology Score (Hilbe, 2008). For example, the effect of treatment on Conser-vative subjects (Ideology score 2) is:

$$IRR_{Treatment \times Ideology_2} = e^{\widehat{\beta_{Treatment}} + \widehat{\beta_I deology} \times 2}$$

Before presenting the results on post-treatment peace tweets, it is important to acknowledge that we also estimated this model using as an outcome variable the post-treatment sentiment score of our subjects, to test if the intervention could cause people to change their expressed opinion. We find null effects for all of our treatments, except for a positive effect in the first day after treatment (results in Appendix A). These Day 1 results primarily consist of replies to the bots, and are driven by people in this short time frame who otherwise not have tweeted about the Peace Process (and have been assigned sentiment 0) tweeting something positive (and being assigned sentiment 1). This is not surprising, given the highly polarized context in which the experiment took place and the low intensity of our intervention.

However, we do find effects on post-treatment tweeting behavior. Initially, we esti-mated a model as described in our Pre-Analysis Plan with all six treatments interacted with the three-level Ideology of the subject. Our models indicated extremely minor variation in treatment effects based on the identity of the sender used to deliver the treatment, or in the message used. Despite our initial hypothesis that the identity of the sender would play a significant role in subjects' response to the treatment, and to ameliorate concern about the problem of multiple comparisons, we collapse the 6 treatments conditions into one.[14]

The results in the following Figures are displayed in four non-overlapping time periods: Week 1, Week 2 and Weeks 3/4. For ease of visualization, the results from Day 1—that is, the first 24 h after treatment, too short a time period for calculating meaningful percentage changes—are not displayed. Week 1 results are hours 25–168 after treatment, and so on. This raises another concern about multiple

[12] Details about the implementation of the NB model can be found in Appendix D.

[13] For a detailed regression analysis, see Appendix B. The decision to use the Negative Binomial model is justified due to the overdispersed data, and the decision to include the logged pre-treatment tweet count variables is due to improved model fit.

[14] The results of the experiment disaggregated by the identity of the sender are presented in Appendix.

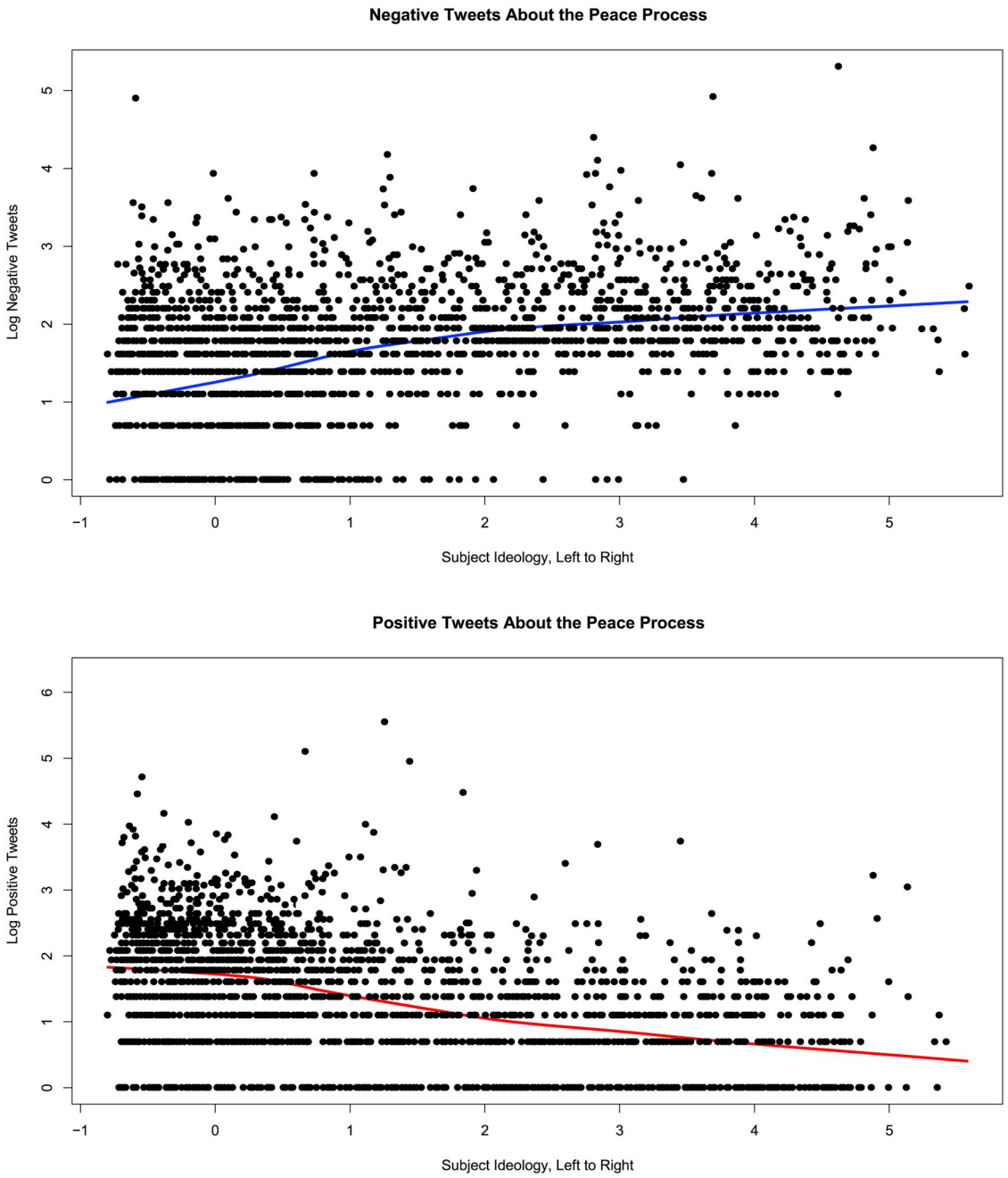

Fig. 4. Validating tweet sentiment and estimated subject ideology.

comparisons: each Figure contains the results of three separate regressions. In Appendix C we show that our results are robust to the most common multiple comparisons corrections—i.e. Bonferroni, Benjamini and Hochberg, and Holm's methods.

Moreover, following Munger (2017b), we expect that (excluding Day 1, in which effects are driven by direct responses to the treatment tweet)[15] effects will only decay. That is, we assume that it if we find a null result in Week 1 and a significant result in Week 2, the latter must be spurious. This idea is consistent with an intuitive model of tweeting behavior: although our treatment may have an effect, whatever processes are causing a given subject to select a baseline level of tweeting are still operating during this time period, causing them to return to that baseline level.[16]

Fig. 5 reports the experimental results on the full sample of 3516 subjects for each of the four time periods. Here, we do not interact treatment with subject ideology.[17] In all of the Figures that follow, the lines can be interpreted as the % change in the number of tweets (either positive or negative, depending on the Figure) the subject sent in the specified time period, relative to a control subject who did not receive

[15] Section E in Appendix presents a thorough analysis of the direct responses to bots.

[16] We point out that the continuous measurement strategy is an asset of our research design. Other approaches (say, that require costly survey responses) might only be able to record one post-treatment outcome measure. The rightmost half of our plots is strictly an advantage of our approach.

[17] We were unable to calculate the ideology of 775 of the 3516 subjects because they did not follow enough political accounts, so we do not include ideology as a covariate in these models. Results are unchanged if we restrict the analysis to only those subjects for whom we have an ideology estimate.

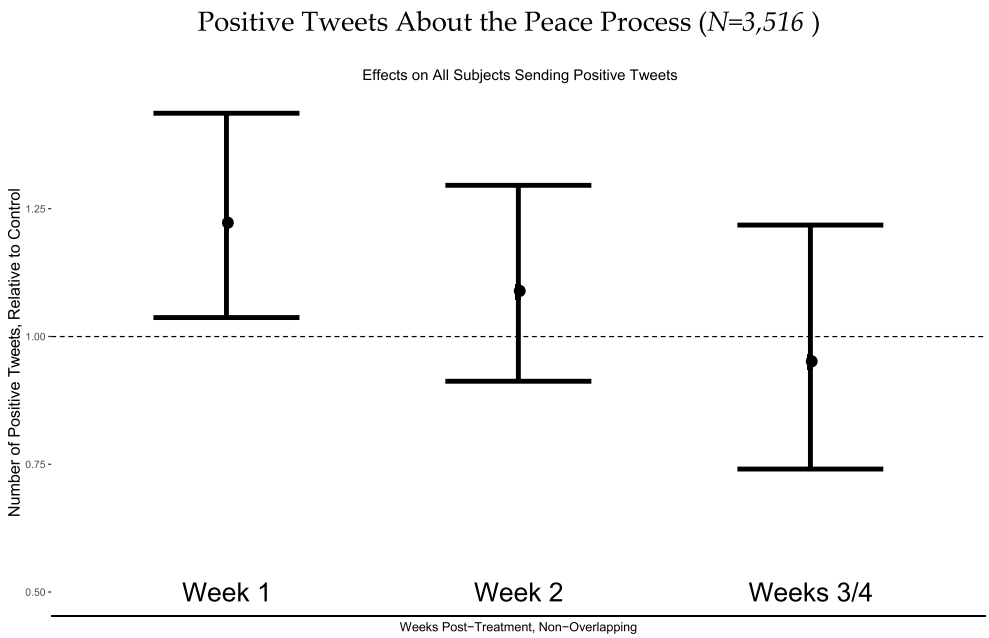

Fig. 5. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in non-overlapping time periods after treatment. For example, the Incidence Ratio of 1.25 in Week 1 indicates that treated subjects sent 125% as many positive tweets about the peace process as untreated subjects. The bars represent 95% confidence intervals.

treatment.

For example, the first line in Fig. 5 ($IRR_{T\ reatment} \approx \mathbf{1.25}$), indicates that treated subjects experienced a positive 25% change in sending positive tweets about the peace process, compared to no change for untreated subjects.[18] These are ratios: going from 0.5 to 1 represents the same effect size (a 100% increase) as going from 1 to 2, so the upper half of the confidence intervals appear longer than the lower half.

Overall, Fig. 5 indicates a small but significant increase in positive tweets sent in Week 1. This effect decays in Week 2, as expected, and is almost exactly 0 in Weeks 3/4.

Fig. 6 breaks up the analysis by the ideology of the subject. Here, we see that the increase in Week 1 is largest among Liberals. Among Conservatives, in fact, we estimate a slight *decrease* in positive tweets. Again, this effect decays in Week 2, and all three estimates are nearly exactly 0 in Weeks 3/4.[19]

We also need to see if our interventions caused any change in the rate of sending *negative* tweets—tweets that argued against the peace process. Figs. 7 and 8 replicate the above results, but for negative tweets.

The un-interacted results in Fig. 7 show that there were no significant increases in sending negative tweets in any time period. There does, however, appear to be a slight increase in the point estimate between Week 1 and Week 2. Fig. 8 reveals this to be driven by small increases among the Conservative and Moderate subject pools, the latter of which indicates a significant effect in Week 2.

Per our assumption of monotonic decay above, we believe that this increase is spuri-ous. There is no plausible mechanism by which moderates would respond to treatment more in days 8–14 after treatment than in days 2–7.

Overall, these results support our argument that conservatives do not feel encour-aged to send more negative tweets, and perhaps—with the exception of those who reply directly to the bots, discussed in detail in Appendix D–simply ignore the messages.

7. Conclusions

We performed a randomized experiment on Twitter users who we identified as interested in the peace process in Colombia. To our knowledge, this is the largest-scale social science Twitter bot experiment conducted to date. We tested several hypotheses about the potential effects that these strategies may have on preferences and attitudes. To do so, ahead of the plebiscite we sent public messages to users encouraging them to support the peace process, varying the identity of the information source and the content. We sent two types of messages, a conservative message and a liberal message, from three different types of bots, namely that of a scientist, a priest and a general, which are respected non-political public figures in Colombia.

Our goal was to learn if public figures and messages more aligned with subjects ideological preconceptions, would be more effective at encouraging people to support and talk more about the peace process. Our results show that we could not cause subjects to change their expressive behavior in favor of the peace process. We are not surprised by these null effects, similar to findings in alternative contexts (Broockman and Green, 2014; Kalla and Broockman, 2017). The peace plebiscite took place in a highly polarized environment and the result of the elections reflects it. Information flows ahead of the election were massive, so that it would be hard to change people's opinions using this type of strategy.

While we find that bots have a limited ability to change expressed opinions, the results should be taken with care. Bail et al. (2018) show that Democrats and Republicans in the U.S. significantly changed their views after following liberal and conservative Twitter bots. One potential explanation for the different results is the sophistication of their

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

[19] The confidence intervals in Fig. 6 are calculated from the estimated variance of this estimator:

$$V_{priest \times Ideo\ \log y_1} = V(\widehat{\beta}_2) + Ideo\ \log y^2 V(\widehat{\beta}_6) + 2Ideo\ \log\ y \times Cov(\widehat{\beta}_2 \widehat{\beta}_6)$$

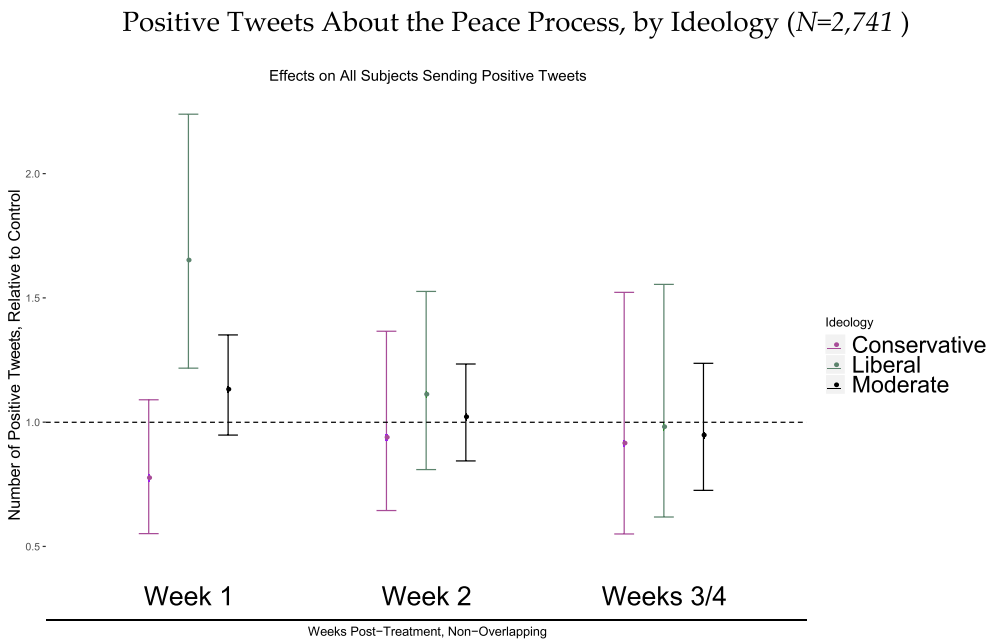

Fig. 6. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in non-overlapping time periods after treatment, interacted with the ideology of the subject. For example, the Incidence Ratio of 1.65 for Liberals (in green) in Week 1 indicates that treated liberal subjects sent 165% as many positive tweets about the peace process as untreated liberal subjects. The bars represent 95% confidence intervals.

bots. More realistic and potentially malicious bots might be more effective in changing opinions or attitudes toward certain topics.

We also have evidence that certain combinations of senders and messages might backfire, as conservatives had differential reactions when approached by a liberal scien-tist or a conservative priest (see Section E of the Online Appendix). Hence, in terms of the moral reframing theory (Feinberg and Miller, 2015; Volkel and Feinberg, 2017), we have learned that not only the content of messages matter, but also the identity of the sender and if it is aligned with the receiver's ideological position.

In many ways, the election studied in this paper is quite unusual. A plebiscite to endorse the agreement signed by a central government and a guerrilla group in a developing country, is a rare event. However, many of the characteristics of this election resemble what has happened—a is going to happen—elsewhere. A deeply polarized society in which social media, elites, and public figures play a key role at shaping citizens' opinions and their subsequent political decisions. In such context, non-political public institutions, like the ones used in this experiment, may increase the debate and get people to talk more, but in highly unexpected ways.

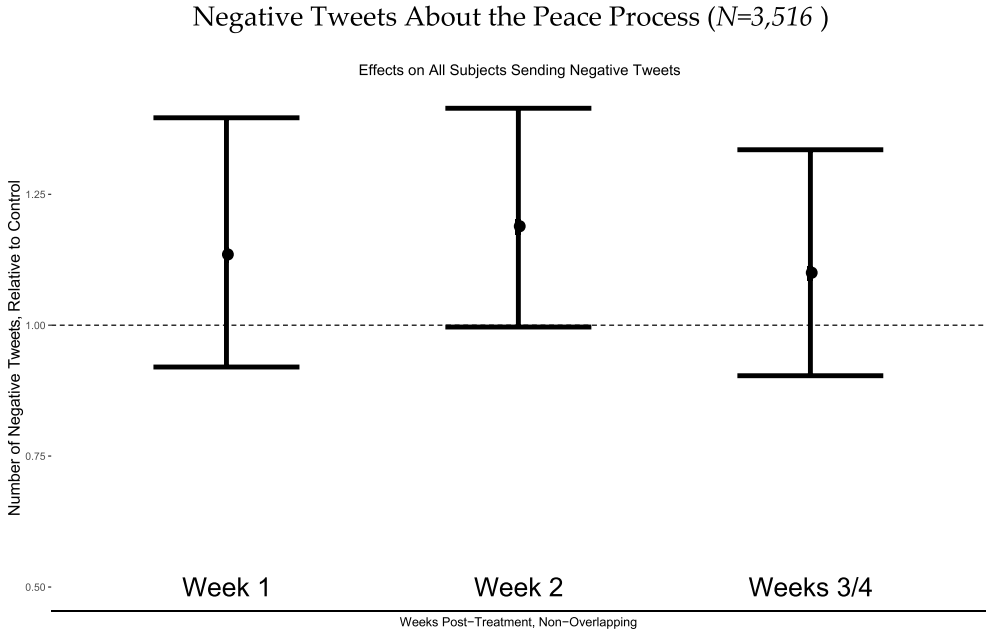

Fig. 7. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in non-overlapping time periods after treatment. For example, the Incidence Ratio of 1.15 in Week 1 indicates that treated subjects sent 115% as many negative tweets about the peace process as untreated subjects. The bars represent 95% confidence intervals.

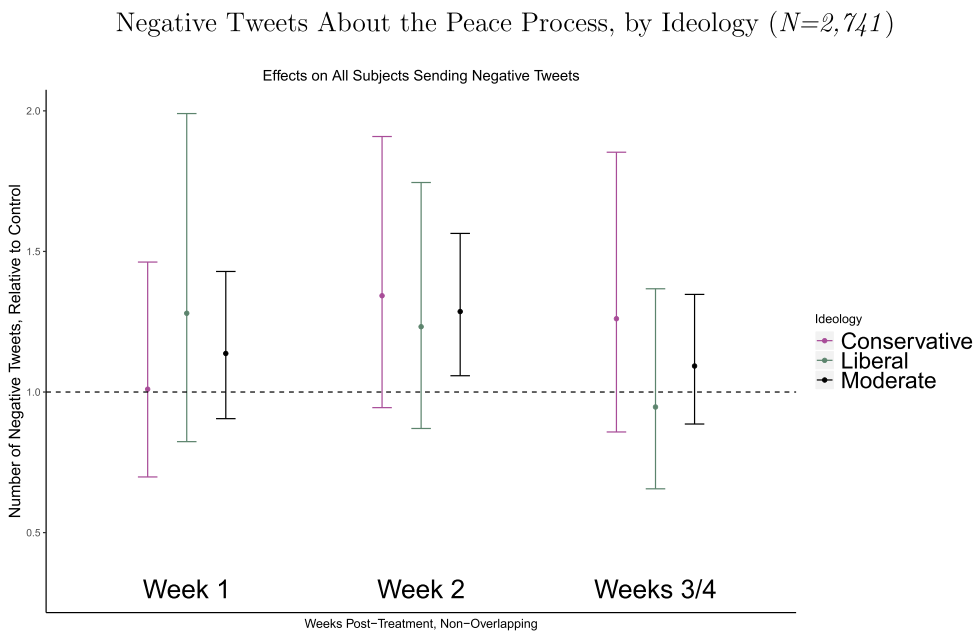

Fig. 8. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in non-overlapping time periods after treatment, interacted with the ideology of the subject. For example, the Incidence Ratio of 1.25 for Liberals (in green) in Week 1 indicates that treated liberal subjects sent 125% as many negative tweets about the peace process as untreated liberal subjects. The bars represent 95% confidence intervals.

Appendix

A Null Effects on Mean Sentiment

To examine changes in the mean sentiment of tweets sent by subjects, we averaged the sentiment of all the tweets they sent in a given time period. Assuming a base sentiment of 0 if they did not tweet about the peace process in that time period, we added the number of positive tweets, subtracted the number of negative tweets, and divided by the total number of tweets. Because this is no longer count data, we estimate the results using OLS. Results are essentially identical when the data is transformed to the [0,1] space and estimated with logistic regression.

Table 1Table 1 shows a significant increase in the average sentiment of subjects in the first day after treatment. Per the discussion in the body of the paper, this is largely driven by positive responses to the treatment tweet, and the modal unit in this short time period is going from sentiment of 0 (no tweets) to a sentiment of 1 (positive tweet). In contrast to the results in Fig. 6, however, there is no effect in the 1 Week or 2 Weeks time periods. There is a marginal negative effect in the 1 Month time period, but again per the discussion in the body of the paper, there is no plausible mechanism for this effect to be so delayed, and this significance is likely spurious.

Table 1
Treatment Effects on Mean Sentiment

	1 Day	1 Week	2 Weeks	1 Month
	(1)	(2)	(3)	(4)
Mean Pre-Treatment Sentiment	0.101*** (0.010)	0.341*** (0.016)	0.173*** (0.017)	0.039** (0.015)
Treatment	0.066** (0.026)	0.008 (0.044)	−0.062 (0.047)	−0.072* (0.042)
Conservative (0, 1 or 2)	−0.007 (0.020)	−0.129*** (0.033)	−0.083** (0.036)	−0.032 (0.032)
Treatment*Conservative	−0.016 (0.021)	−0.018 (0.036)	−0.010 (0.038)	−0.035 (0.034)
Constant	0.040 (0.024)	0.368*** (0.041)	0.187*** (0.043)	−0.028 (0.039)
Observations	2741	2741	2741	2741
R^2	0.066	0.252	0.081	0.004
Adjusted R^2	0.064	0.250	0.080	0.003

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

The alternative model for sentiment would drop the assumption of a sentiment score of 0 for people who did not tweet during a given time period. The results of this analysis can be found in Table 2Table 2. Here there are no signficant treatment effects.

Table 2
Treatment Effects on Mean Sentiment—Drop Non-Tweeters Time Period (Non-Overlapping):

	1 Day	1 Week	2 Weeks	1 Month
	(1)	(2)	(3)	(4)
Mean Pre-Treatment Sentiment	0.543*** (0.070)	0.567*** (0.029)	0.392*** (0.036)	0.176*** (0.049)
Treatment	−0.065 (0.175)	0.004 (0.071)	−0.063 (0.088)	−0.188 (0.122)
Conservative	0.505***	0.279***	0.135*	0.063
Conservative	−0.505*** (0.167)	−0.279*** (0.060)	−0.135* (0.072)	−0.063 (0.092)
Treatment*Conservative	0.275 (0.178)	−0.028 (0.065)	−0.063 (0.078)	0.068 (0.099)
Constant	0.617*** (0.167)	0.580*** (0.067)	0.246*** (0.082)	−0.128 (0.113)
Observations	324	1413	1330	880
R^2	0.271	0.391	0.167	0.023
Adjusted R^2	0.261	0.390	0.164	0.019

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

B Modelling Decisions

Table 3 Table 3displays the detailed results of the models displayed in the body of the paper. The time period in the table is the critical one identified in 6: the first week after treatment for positive tweets. Two models are considered: the left column uses logged pre-treatment tweet counts, while the right does not. The former is preferred on model fit grounds. Both models, though, report θ significantly below 1. θ is the parame-ter calculated by the Negative Binomial model fit with the *R* package "MASS" (Choi et al., 2014). This parameter is the inverse of the α parameter sometimes used to rep-resent overdispersion (the typical interpretation of which is that values over 1 represent overdispersion).

Table 3
Change in Positive Tweets One Week Post-Treatment

	Logged Tweet Counts	Unlogged Tweet Counts
Negative Pre-Treatment Tweets	0.554*** (0.040)	0.028*** (0.005)
Positive Pre-Treatment Tweets	0.888*** (0.033)	0.038*** (0.002)
Treatment	0.502*** (0.156)	0.498*** (0.164)
Conservative (0, 1 or 2)	−0.021 (0.127)	−0.419*** (0.128)
Treatment*Conservative	0.378*** (0.138)	0.369*** (0.141)
Constant	0.539*** (0.167)	0.452*** (0.153)
Observations	2741	2741
Log Likelihood	4260.654	4422.493
θ (inverse dispersion)	0.422*** (0.019)	0.335*** (0.015)
Akaike Inf. Crit.	8533.307	8856.986

Note: *p < 0.1; **p < 0.05; ***p < 0.01.

C Robustness to Multiple Comparisons Corrections

In this section, we report the results of a series of multiple comparisons adjustments that we implement on our main estimations. For this purpose, we use the Bonferroni, Benjamini and Hochberg, and Holm corrections. In each case, we determine the num-ber of hypotheses that we want to test and adjust the p-values in accordance with that number. For our study, the coefficient of interest is the interaction between our *Conservative* measure (which takes values of 0, 1, or 2) and the treatment indicator (0 or 1). Remember that the latter has been aggregated, in such a way that it equals 1 for subjects that receive any form of treatment.

We have six main outcomes of interest, and accordingly, six hypotheses to test: the (logged) count of positive tweets on week 1, week 2, and weeks 3 and 4, and the (logged) count of negative tweets on week 1, week 2, and weeks 3 and 4. We want to deter-mine if the effect of receiving any treatment on each of these six outcomes, is mediated by the ideology measure. Consequently, we focus on the p-value of the coefficient of *Treatment * Conservative*. Table 4Table 4 reports the results of the three adjustments, for each of the six coefficients of interest (one per outcome). As it is clear from this Table, no matter if we use Bonferroni, Benjamini and Hochberg, or Holm's method, the only coefficient that is always significant, is the term associated to the regression on positive tweets during week 1. Remember that this coefficient is negative. Hence, the main result of this paper—that liberal subjects tweet more in favor of the peace process after being treated and during the first post-treatment week—is robust to the most common multiple comparisons corrections.

Table 4
Multiple Comparisons Corrections

Positive Tweets Week 1	0.006	0.037	Yes	1	0.008	Yes	0.008	Yes
Positive Tweets Week 2	0.568	3.410	No	4	0.033	No	0.017	No
Positive Tweets Weeks 3/4	0.866	5.194	No	6	0.050	No	0.050	No
Negative Tweets Week 1	0.493	2.956	No	3	0.025	No	0.013	No
Negative Tweets Week 2	0.771	4.629	No	5	0.042	No	0.025	No
Negative Tweets Weeks 3/4	0.369	2.211	No	2	0.017	No	0.010	No

Treatment*Conservative p-value Bonf p-val Bonf sign Rank BH Critical Val BH Sign Holm Critical Val Holm Sign.

D Details of constructing the sentiment classifier

The data collection process was a fundamental to the experiment; all of the data are from Twitter. Twitter grants access to the public interested in its data through an API which can be used to scrape and organize tweets from a user, a hashtag, a topic or even a location. The methodology of data collection was focused on fetching data from Twitter based on a shifting set of words and n-grams related to the Peace Process in Colombia.

We updated the search terms over the course of the eight months before the peace process, using terms that related to current developments. These terms were filled into Twitter's standard search API as search terms. Then, the Twitter API returned the list of tweets that mentioned any of the search terms in the previous 7 days, subject to an individual-account cap on the limit in the number of tweets the user is allowed to download. We programmed an algorithm that repeated this process automatically every hour, stacking the tweets in a database. This first stage ran from March 2016 to October 2016, and resulted in 1 million tweets regarding the peace process in Colombia. This database contains not only the text of every tweet, but also the name of the user, the Twitter user name, the number of retweets, the number of favorites, the self-reported location, the geocoded location (not always available), the biography of the user, and other variables.

This database was used afterwards to identify the users who shared at least five tweets in our initial collection process. The information was filtered in terms of number of tweets, number of followers, location and other characteristics of the accounts. We ultimately selected 4500 non-institutional and non-bot users with fewer than 2000 followers who were active in terms of sharing comments about the peace process from July to September of 2016 and located in Colombia.[20]

Fig. 9Fig. 9 describes our sample selection procedure. From the initial pool of 4500 subjects, we restricted our analysis to the 3516 who sent at least five pre-treatment tweets containing one of the 33 terms defined by the dictionary in Footnote 11. This restricted dictionary was necessary to ensure the comparability of our definition of what.

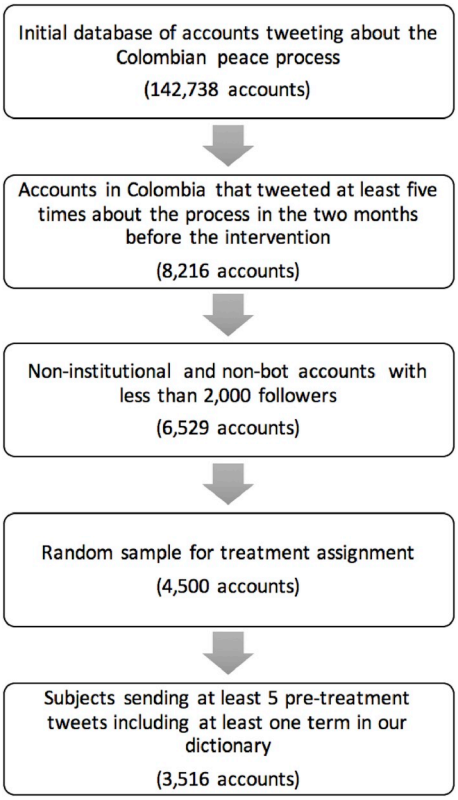

Fig. 9. Sample Selection Process

This flowchart depicts the selection process of the accounts that were ultimately used in the experiment. makes a tweet "about the peace process" between the pre- and post-treatment periods. The "attrition" in this stage was not statistically significantly related to treatment, so the inference from randomization is not threatened. All of the "attrition" was due to our mistaken inclusion of subjects who did not fit our pre-treatment sampling frame due to the time constraint of completing the experiment before the peace process.

The second purpose of the database was to use labeled tweets as examples to train a machine learning model able to tell the sentiment of a tweet in the context of the peace process of Colombia. This process of building the machine learning classifier had four main steps: Data cleaning, data labeling, feature extraction and selection, training the algorithm and calibration of parameters. The final product is a method to calculate the sentiment score of any tweet from negative to positive. The first step involves making the text of each tweet readable for a computer, this means taking away uncommon

[20] We used intuition and common sense to identify institutional accounts and bots, as the automatic bot detection tools available at the time did not perform well for accounts tweeting in Spanish. While we acknowledge the possibility that some bots ended up in our subject pool, our random assignment of subjects to treatment suggests that this is not a problem for inference. Furthermore, we expect that bots should be "never-compliers" in the sense that they never respond to our treatments, and their presence should thus bias against finding significant treatment effects.

symbols, accents, icons, upper case letters and extra spaces. After this, we identify a set of words called stop words inside every tweet and delete them, this is necessary given that not every word contributes information about the sentiment of a tweets, for example: an, any, or, to, the. The third step consists on performing stemming to the words of each tweet. This process seeks to collapse the words with same meaning, but different conjugations, for instance, the word negotiating has the same base meaning as the word negotiated, therefore it would be useful if the computer understands these two words as the same one. In this case, both terms would be converted to its base word or stem, which in this case corresponds to the word negotiate.

The second phase is responsible for the development of a set of examples whose main purpose is to teach the machine learning model to classify correctly. We manually labeled a random set of tweets according its sentiment towards the peace process (positive or negative). After preliminary tests about out-of-sample accuracy, we settled upon using the Naive Bayes binary classifier.

Before the training stage, the text present inside every tweet needs to be expressed in a structured form (units of observation with a set of characteristics expressed as rows and columns). Our approach is to use the method bag of words, expressing words inside a tweet as binary variables. In this sense, every tweet represents a row of the data frame with as many variables as possible words in a tweet. This means that the number of variables depends on the size of the vocabulary present in the corpus used. At the end of this process we had every tweet expressed as a row of zeros and ones (one if the word appears in the tweet and zero if not).

Finally, having all the text structured in a database, we trained a Naive Bayes binary classifier with the set of labeled tweets. The basic premise of this algorithm is to use the words present in a tweet to estimate the probability that its sentiment is positive or negative.

$$\widehat{c} = argmaxP(c|d) \tag{1}$$

Essentially, the intuition behind the use of Bayes theorem to classify text is to simplify equation (1) and make a naive assumption regarding the interaction between the words inside a document. In equation (1) $\hat{c}$ expresses the estimated class c given its probability. Having the Bayes theorem expressed in equation (2) with d as the document and c as the class (in our case either positive or negative), we can substitute equation (1) into equation (2) to have the expression (3).

$$P\left(c|d\right) = \frac{P(d|c)P(c)}{P(d)} \tag{2}$$

$$argmaxP\left(c|d\right) = \frac{P(d|c)P(c)}{P(d)} \tag{3}$$

In this sense, since the marginal probability $P(d)$ is equal for all classes, it can be disregarded of equation (3) and we can simplify the equation to:

$$argmaxP(c|d) = P(d|c)P(\mathrm{c}) \tag{4}$$

Then, the probability of class c is given by the multiplication of the prior probability of the class c and the likelihood of the document d given the class c. After this, the document can be represented as a set of features (in this case words). At this point, we assume that the words of a document are independent from each other. This let us express the likelihood as the multiplication of the probabilities of every single word in the document (words expressed as w). Therefore, we are calculating for each word, in a set of a labeled documents, the probability it appears given the class of the document, and then these results are multiplied by the prior probability of the class c as seen in equation (5). In this case, the training database is used to calculate both the prior probability of each class and the conditional probability of every single word inside our corpus.

$$\widehat{c} = argmaxP(\mathrm{c})\Pi P(w|c) \tag{5}$$

We used cross validation to evaluate the results and the performance of the classi-fication algorithm. This means that we randomly divided our labeled database in 10 equally sized folds, and then for each one we calculated and evaluated our Naive Bayes classifier. For the purpose evaluating the results, each fold is divided into a training sample and a test sample. The test samples provide the ability to compare the true classes for every tweet versus the predicted ones. At the end, to compute the overall precision, a simple average between folds is calculated.

E Additional Results: Direct Responses

We now analyze the direct responses to the bots' tweets. For this purpose, keeping constant the identity of the bot, we estimate the effect of a liberal message—versus a conservative one—on the probability of reacting to the direct mention made to each account. Formally, for each bot k, where $k =$ General, Priest, Scientist, we estimate models of the type:

$$Reaction_{\mathrm{i}} = \beta_{\mathrm{k0}} + \beta_{\mathrm{k1}} Liberal\_Message_{\mathrm{i}} + \varepsilon_{\mathrm{i}}$$

where $Reaction_i$ is a dummy variable indicating whether subject i reacts to the message sent by the bot or not, $Liberal\_Message_i$ is a dummy variable that indicates if subject i received a liberal message from bot k, and ε_i is the error term. We estimate separate models for any type of reaction, as well as for exclusively positive or negative reactions. Positive reactions correspond to likes, retweets, or positive replies to the bot. On the other hand, negative reactions are associated with negative replies.[21] The coefficients of interest in this set of regressions are β_{k1}. If this coefficient is positive for bot k, it

[21] Manual coding for these replies was performed, to determine whether the subject responded posi-tively or negatively to the bot.

means that subjects tweeted by such bot tend to respond more (positively or negatively) when the message has a liberal content, as compared to the conservative message.

Keeping constant the identity of the bot, these coefficients indicate if liberal or conservative messages produce more reactions.

Fig. 10Fig. 10 plots the regression coefficients—and the associated confidence intervals— of these models for the liberal versus conservative versions of each of the three types of bots. Each of the three outcomes in the Figure (any reaction, positive reaction and negative reaction) are the result of a separate OLS regression—results are substantively the same if a logit model is used instead. Values above 0 in Fig. 10 mean that outcome was more likely to be caused by the liberal message that type of bot, while values below 0 indicate higher likelihood for the conservative message.

The results in Fig. 10 indicate that the liberal general caused more positive reac-tions than the conservative general, and that the liberal scientist caused fewer positive reactions and more negative reactions. In both cases, then, the bots that sent messages "against type" (liberal messages sent by the general and conservative messages sent by the scientist) were more likely to engender positive reactions than messages "with type." As we expected, there were no differential effects of the liberal priest compared to the conservative priest. In order to understand the channels driving these results, we disaggregate the effects of these messages along the ideology dimension: we test whether there are differential effects for liberal, moderate, and conservative subjects.

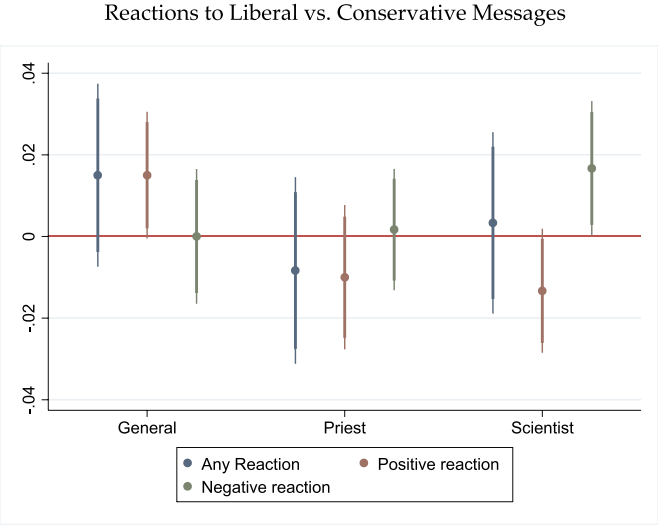

Fig. 10. Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot, while values below 0 indicate higher likelihood for the conservative version.

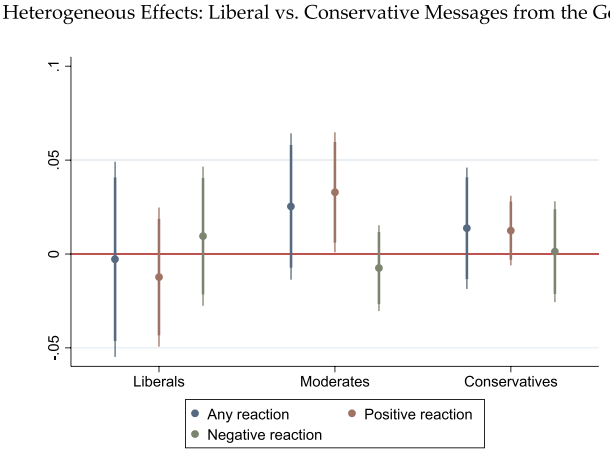

Fig. 11. Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative version.

The results in Figs. 11, 12, and 13Figs. 11, 12 and 13 represent heterogeneous effects at the ideology level. These results reflect that the positive effects of liberal messages sent by the General are mainly driven by moderate subjects (Fig. 11). Additionally, the increase in negative reactions to liberal messages sent by the scientist are driven by conservative subjects disliking these messages (Fig. 13). Finally, in the case of the priest, conservative subjects are more likely to react positively when they receive a conservative message from this type of bot. Note that in some cases there is no

point estimate. This occurs when there is no variation in the outcome variable. For example, in the case of liberals who received a message from the scientist, none of them had a negative reaction.

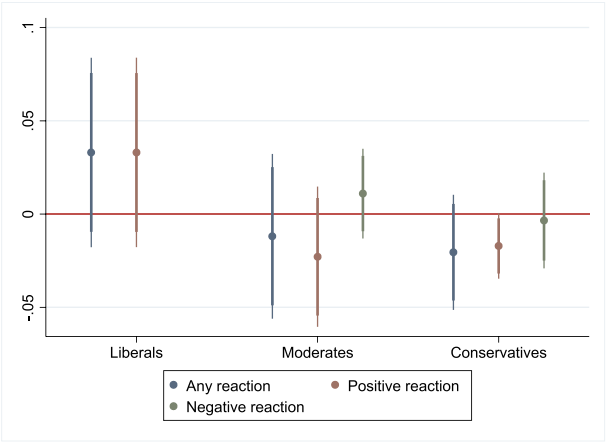

Fig. 12. Each of the three outcomes are the result of a separate OLS regression. Values above 0 in Figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative version.

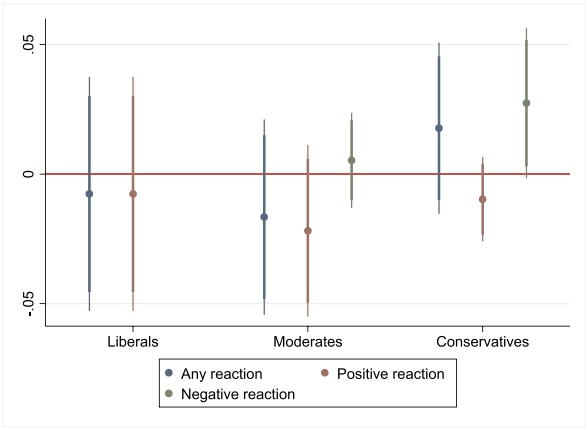

Fig. 13. Each of the three outcomes are the result of a separate OLS regression. Values above 0 in the figure mean that outcome was more likely to be caused by a message from the liberal version of that type of bot for each subgroup of subjects, while values below 0 indicate higher likelihood for the conservative version.

Overall, these results reveal that in general conservative subjects are the ones who react differentially to these messages and tend to respond directly to our bots, especially depending if the message is "against" of "with" type. They dislike liberal messages sent by the scientist–relative to conservative messages from same—and are more likely to react in a positive way when contacted by the conservative priest compared to the liberal one. This goes in line with theoretical concepts outlined above, such as the 'perceptual screen' theory (Campbell, 1960), in the sense that a subject's moral values influence the reaction they have to our encouragement (Kernell and Mullinix, 2013). The novelty of our approach is that it allows us to show that the *identity* of the sender, and not only the *content* of the message, matters.

F Disaggregated analysis of treatment effects by identity of sender

The experimental results on the sample of liberal subjects disaggregated by the type of sender are displayed in Fig. 14Fig. 14. As in the main text, the dependent variable is the number of tweets (either positive or negative) the subject sent in the specified time period. We exclude the first day in which there were many direct reactions to the tweets.

$IRR_{scientist \times Ideo\ \log y_0} \approx 1.5$, the effect of the conservative scientist treatment on liberal subjects during the first week after the intervention, can be seen in the black line in the left section of the plot. This Incidence Ratio implies that the average subject with Ideology Score 0 (liberal) who received the conservative scientist treatment tweeted about 150% as many positive tweets about the peace process as the average subject with Ideology Score 0 in the control condition.[22] The confidence intervals in Fig. 14 are calculated from the estimated variance of this estimator:

$$V_{priest \times Ideology_1} = V(\hat{\beta}_2) + Ideology^2 V(\hat{\beta}_6) + 2Ideology \times Cov(\hat{\beta}_2 \hat{\beta}_6)$$

These are ratios: going from 0.5 to 1 represents the same effect size (a 100% increase) as going from 1 to 2, so the upper half of the confidence intervals appear longer than the lower half. Also, recall that the Liberal and Conservative samples each comprise 25% of the overall sample compared to 50% for the moderate sample. Because the sample is twice as big, the standard errors for the moderate sample are smaller.

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

In general, all six of the treatment conditions had similar effects on liberals' rate of sending positive tweets: Liberal respondents were encouraged to send more positive tweets during the first week after treatment. As can be seen on the right hand portion of this figure, this effect disappears after the first week.

Fig. 15Fig. 15 shows that our treatment conditions have no effect, in general, on the subsample of moderate subjects. In the case of conservative users, as shown in Fig. 16Fig. 16, for the first week the point estimates are negative but non-significant. In sum, our treatments encourage liberals to tweet more about the peace process during the first week, but no effects are produced on conservatives or moderates at any given time.

We also need to see if our interventions caused any change in the rate of sending.

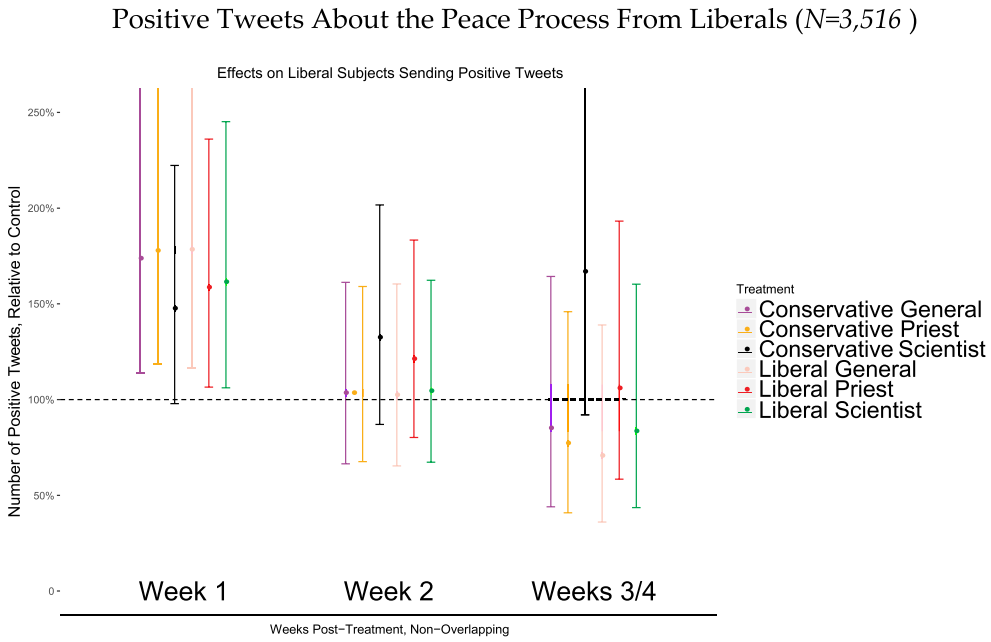

Fig. 14. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, exclud-ing day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 130% as many positive tweets about the peace process as the subjects with Ideology Score 1 in the control group. The bars represent 95% confidence intervals.

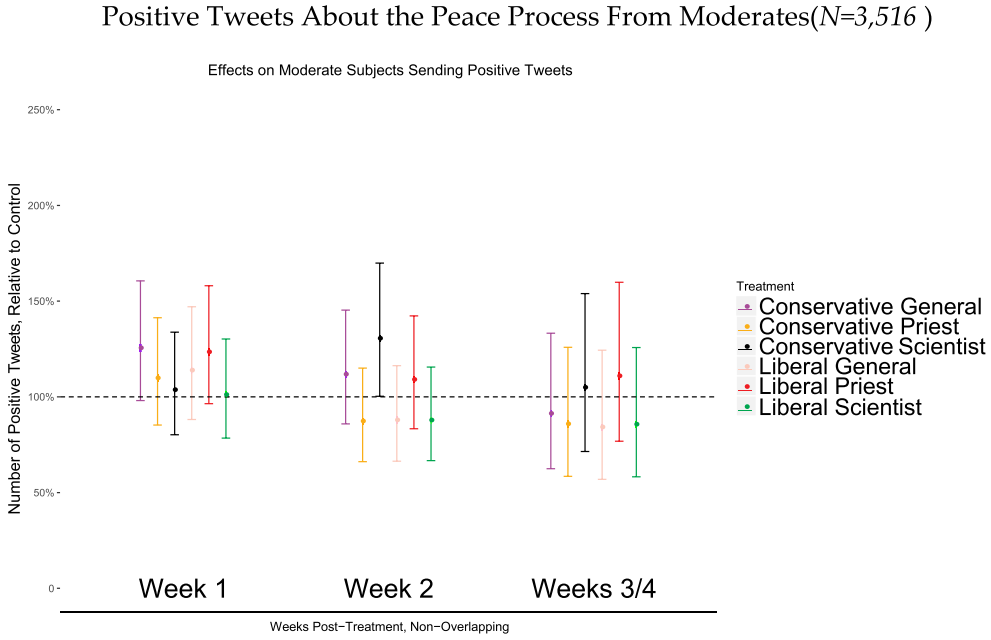

Fig. 15. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, exclud-ing day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the

plot means that these subjects sent 130% as many positive tweets about the peace process as the subjects with Ideology Score 1 in the control group. The bars represent 95% confidence intervals.

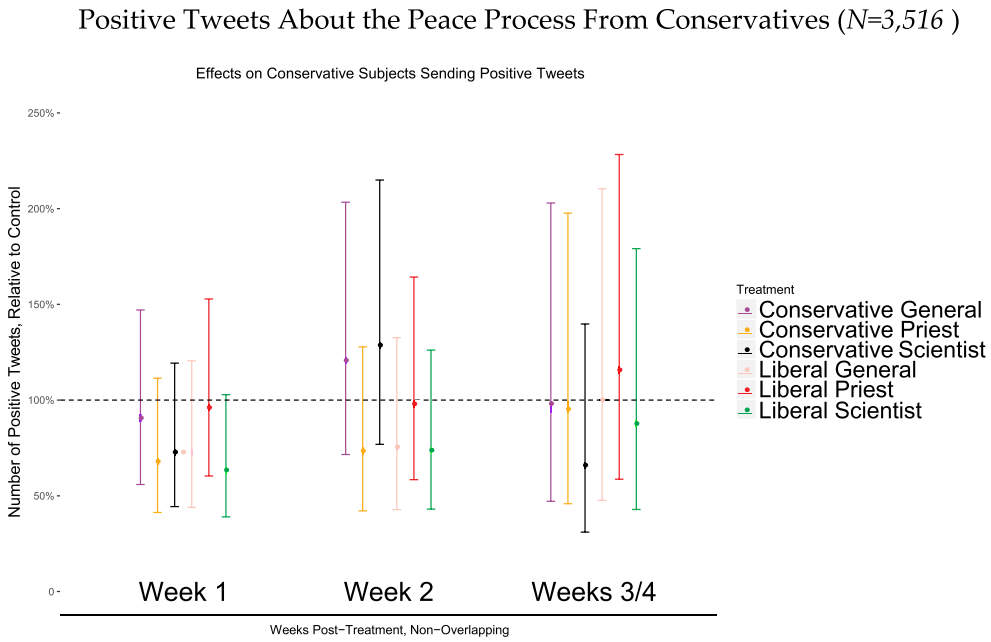

Fig. 16. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the first week after treatment, exclud-ing day 1. For example, the Incidence Ratio of 1.3 associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 130% as many positive tweets about the peace process as the subjects with Ideology Score 1 in the control group. The bars represent 95% confidence intervals. negative tweets—tweets that argued against the peace process. Figs. 17–19Figs. 17, 18 and 19 plot those results.

Encouragingly, across all treatment conditions interacted with subject ideology and the post-intervention week analyzed, only three showed a statistically significant increase in the rate of sending negative tweets about the peace process—precisely the number that we would expect to see by chance. Again, these results support our ar-gument that conservatives do not feel encouraged to send more negative tweets, and perhaps—with the exception of those who reply directly to the bots–simply ignore the messages. *Overall, we do find evidence that bots were able to increase public talk in favor of the peace process, but only among those that to begin with were aligned with the position of the bot and independent of its identity and moral values employed.* After confirming their original preconception, they simply tweet more about it. Hence, these results provide evidence of the so-called "confirmation bias" that characterizes social media.

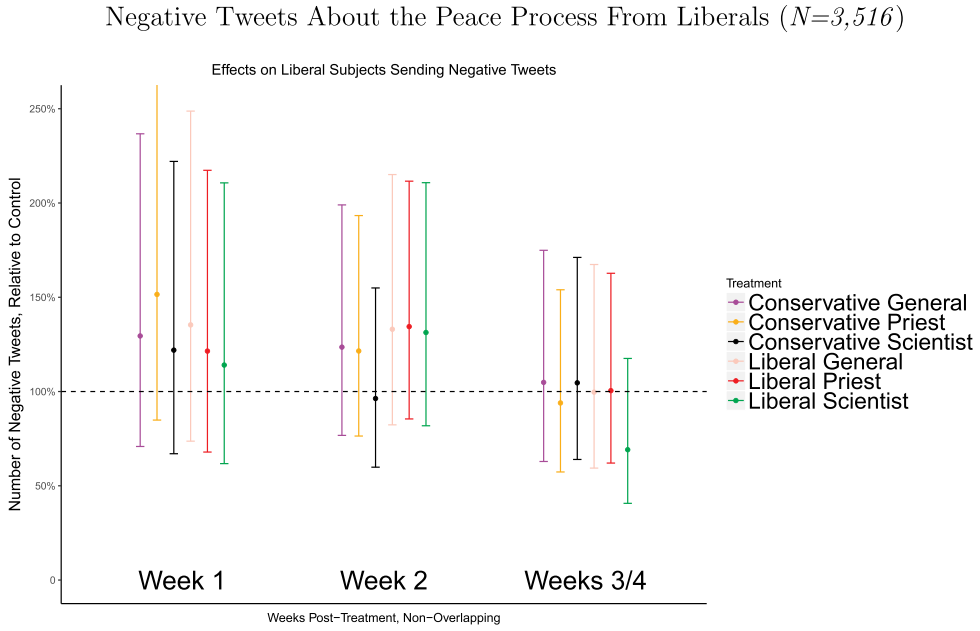

Fig. 17. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot

means that these subjects sent 140% as many tweets about the peace process as the subjects with Ideology Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

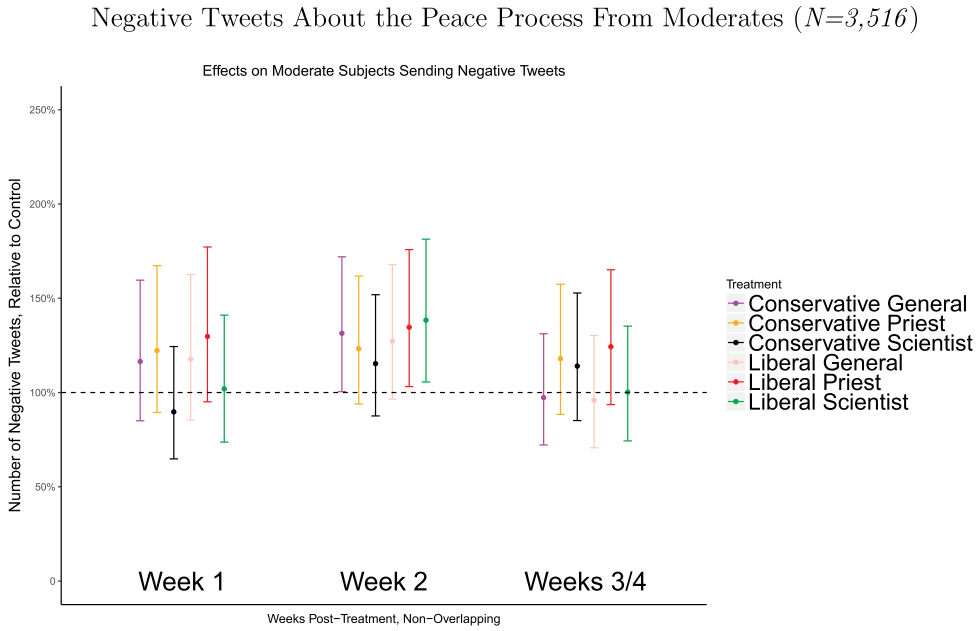

Fig. 18. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 140% as many tweets about the peace process as the subjects with Ideology Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

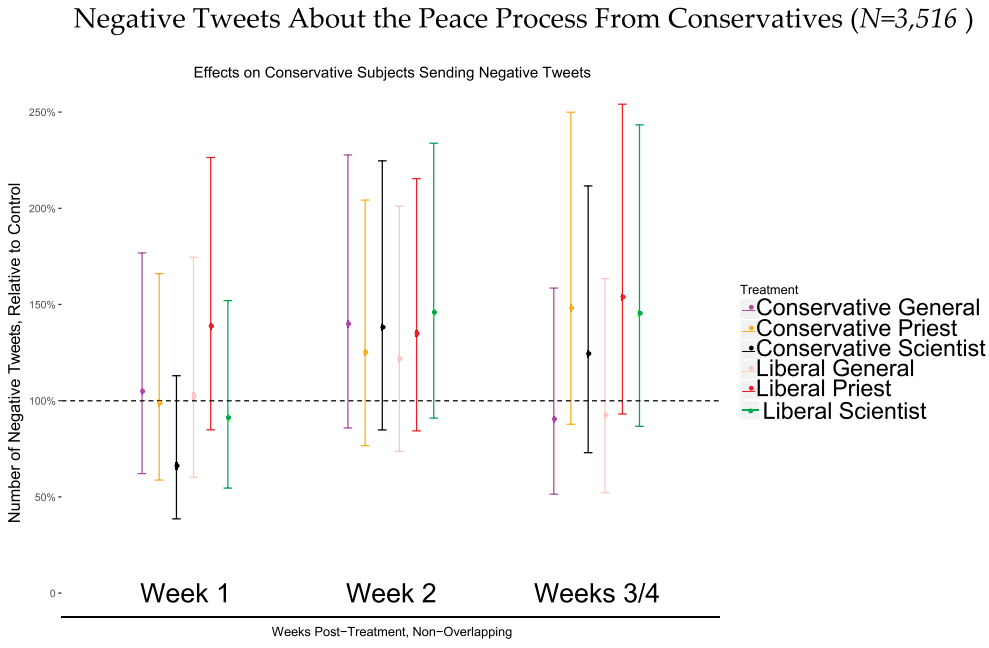

Fig. 19. The Incidence Ratio calculated from the estimated coefficients and variance-covariance matrix from the negative binomial model in the time period from 2 to 7 days after treatment. For example, the Incidence Ratio associated with the liberal priest treatment on subjects with Ideology Score 1 in the middle of the plot means that these subjects sent 140% as many tweets about the peace process as the subjects with Ideology Score 1 in the control group. The thick bars represent 90% confidence intervals and the thin lines represent 95% confidence intervals.

G Ideology and Trust in institutions in the Colombian Context

We presented our bots as representative of people or institutions trusted by Colombian citizens; some of these identities are associated with conservative values, and others with liberal values. According to a Gallop poll conducted in October of 2016, about 60% of the respondents had a favorable opinion of the Catholic Church. This is relatively high, compared to other institutions. The same poll reveals that 71% of the population has a favorable opinion of the military, whose reputation can be explained by their role in the longstanding civil conflict with the FARC. Even though the military tend to be associated with values linked to conservative parties, in this case the perception is at least ambiguous, as the soldiers killed by this war came from cross-cutting segments of society, generating sentiments of sympathy and acknowledgment among all Colombians. Finally, even though opinion polls usually do not ask about citizens' perceptions of scientists, it is well known that academics tend to be more inclined towards liberal values. In fact, in the context of this plebiscite, several of the most renowned Colombian professors signed petitions supporting the peace deal.

In fact, a recent study (British Council, 2017) reveals that the most trusted institu-tions by the Colombian youth are: professors (54%), the army (48%), and the Catholic Church (45%), in sharp contrast with their perceptions of the government, political parties, and illegal armed groups (see Fig. 20). Consequently, we consider that in the Colombian context public figures and institutions associated with the church, the military, and academia, exert influence on citizens' political opinions. We theorized that the priest would be most associated with conservative values and thus the "No" vote and the scientist with liberal values and thus the "Yes" vote, while the soldier would be more moderate.

Opinion polls like the ones conducted by Gallup reveal that the Catholic Church and the army are among the most trusted institutions and public figures in Colombia. These surveys usually do not include respondents' opinions about professors or scientists in general. However, a recent study conducted on young Colombians (British Council, 2017), reveals that the most trusted institutions, in that order, are professors, the army, and the Catholic Church. These are precisely the public figures used in our experiment and the clear motivation on relying on characters that might influence citizens' opinions and behavior.

Fig. 20. Most and Least Trusted Institutions Among Young ColombiansSource British Council (2017). Each percentage represents the proportion of respon-dents that acknowledge trusting each figure or institution.

References

Aggarwal, Anupama, Kumaraguru, Ponnurangam, 2015. What they do in shad- ows: twitter underground follower market. In: 13th Annual Conference on Privacy, Security and Trust (PST). IEEE, pp. 93–100.

Alandete, Davide, 2017. Russian Meddling Machine Sets Sights on Catalonia. *El Pais*.

Bail, Christopher, Argyle, Lisa, Taylor, Brown, Bumpus, John, Haohan, Chen, Hunzakerd, Fallin, Leea, Jaemin, Manna, Marcus, Merhout, Friedolin, Vol- fovsky, Alexander, 2018. Exposure to opposing views on social media can increase political polarization. Proc. Natl. Acad. Sci. 115 (37), 9216–9221.

Barberá, Pablo, 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Political Anal. 23 (1), 76–91.

Bessi, Alessandro, Ferrara, Emilio, 2016. Social bots distort the 2016 U.S. Presi- dential election online discussion. Clin. Hemorheol. and Microcirc. 21 (11).

Black, Laura, Howard, Welser, Cosley, Dan, DeGroot, Jocelyn, 2011. Self- governance through group discussion in Wikipedia: measuring deliberation in online groups. Small Group Res. 42 (5), 595–634.

Brachten, Florian, Stieglitz, Stefan, Hofeditz, Lennart, Kloppenborg, Katharina, Reimann, Annette, 2017. Strategies and influence of social bots in a 2017 German state election - a case study on twitter. In: Proceedings of the Australasian Conference on Information Systems.

British Council, 2017. "Next Generation: aplificando la Voz de los Jóvenes en Colombia." Reporte Preliminar. Retrieved from. https://www.britishcouncil.co/events/lan zamiento-preliminar-investigacion-next- generation-colombia.

Broockman, David, Green, Donald, 2014. "Do online advertisements increase political candidates' name recognition or favorability? Evidence from randomized field experiments. Political Behav. 36 (2), 263–289.

Buckels, Erin, E., Paul, D Trapnell, Paulhus, Delroy L., 2014. Trolls just want to have fun. Personal. Individ. Differ. 67, 97–102.

Campbell, Angus, 1960. The American Voter. University of Chicago Press.

Choi, Meena, Chang, Ching-Yun, Clough, Timothy, Broudy, Daniel, Killeen, Trevor, MacLean, Bren- dan, Vitek, Olga, 2014. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics 30 (17), 2524–2526.

Chu, Zi, Gianvecchio, Steven, Wang, Haining, Jajodia, Sushil, 2012. Detecting automation of twitter accounts: are you a human, bot, or cyborg? IEEE Trans. Dependable Secure Comput. 9 (6), 811–824.

Coleman, Stephen, Blumer, Jay, 2009. The Internet and Democratic Citizenship: Theory, Practice and Policy. Cambridge University Press.

Desposato, Scott, 2015. Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals, vol. 1. Routledge.

Desposato, Scott, 2016. Subjects' and Scholars' Views on Experimental Political Sci- Ence.

Eveland, William, 2004. The effect of political discussion in producing informed citizens: the roles of information, motivation, and elaboration. Political Commun. 21 (2), 177–193.

Feinberg, Matthew, Miller, Robb, 2015. From gulf to bridge: when do moral arguments facilitate political influence. Personal. Soc. Psychol. Bull. 41 (12), 1665–1681.

Ferrara, Emilio, 2017. Disinformation and Social Bot Operations in the Run up to the 2017 French Presidential Election arXiv:1707.00086.

Ferrara, Emilio, Varol, Onur, Davis, Clayton, Menczer, Filippo, Flam- mini, Alessandro, 2016. The rise of social bots. Commun. ACM 59, 96–104.

Graham, Jesse, Haidt, Jonathan, Nosek, Brian, 2009. Liberals and conservatives rely on different sets of moral foundations. J. Personal. Soc. Psychol. 96 (5), 1029–1046.

Graham, Todd, 2015. Everyday Political Talk in the Internet-Based Public Sphere. Edward Elgar Publishing (chapter 14).

Haenschen, Katherine, 2016. Social pressure on social media: using Facebook status updates to increase voter turnout. J. Commun. 66 (4), 542–563.

Halpern, Daniel, Gibbs, Jennifer, 2013. Social media as a catalyst for online deliber- ation? Exploring the affordances of Facebook and YouTube for political expression. Comput. Hum. Behav. 29 (3), 1159–1168.

Hilbe, Joseph M., 2008. Brief overview on interpreting count model risk ratios: an addendum to negative binomial regression. In: An Addendum to Negative Binomial Regression. Cambridge University Press.

Howard, Philip, Kollanyi, Bence, 2016. Bots, #StrongerIn, #Brexit: Computa- Tional Propaganda during the UK-EU Referendum arXiv:1606.06356 [physics].

Johnston, Pamela, Conover, Donald, Searing, Crewe, Ivor, 2001. The deliberative potential of political discussion. Br. J. Polit. Sci. 32 (1), 21–62.

Kalla, J., Broockman, D., 2017. The minimal persuasive effects of campaign contact in general elections: evidence from 49 field experiments. Am. Pol. Sci. Rev. 112 (1).

Kalla, J., Broockman, D., 2018. The minimal persuasive effects of cam-paign contact in general elections: evidence from 49 field experiments. Am. Pol. Sci. Rev. 112 (1), 148–166.

Kernell, Georgia, Mullinix, Kevin, 2013. The Scope of the Partisan 'Perceptual Screen'. Working Paper, retrieved from. https://www.ipr.northwestern. edu/publications/docs/workingpapers/2013/IPR- WP-13-15-A.pdf.

Kim, Joohan, Wyatt, Robert, Katz, Elihu, 1999. News, talk, opinion, participa- tion: the Part Played by conversation in deliberative democracy. Political Commun. 16 (4), 361–385.

King, Gary, Pan, Jennifer, Roberts, Margaret, 2014. Reverse-engineering censorship in China: randomized experimentation and participant observation. Science 345 (6199), 1251722.

Lakoff, George, 2002. Moral Politics: How Liberals and Conservatives Think. University of Chicago Press.

Morgan, Scott, Skitka, Linda, Wisneski, Daniel, 2010. Moral and religious con- victions and intentions to vote in the 2008 presidential election. Anal. Soc. Issues Public Policy 10 (1), 307–320.

Munger, Kevin, 2017. Experimentally reducing partisan incivility on twitter. Working Paper, retrieved from. http://kmunger.github.io/pdfs/jmp.pdf.

Munger, Kevin, 2017. Tweetment effects on the tweeted: experimentally reducing racist harassment. Political Behav. 39 (3), 629–649.

Murthy, Dhiraj, Powell, Alison, Tinati, Ramine, Anstead, Nick, Carr, Leslie, Hal- ford, Susan, Weal, Mark, 2016. Bots and political influence: a sociotechnical in- vestigation of social network capital. Int. J. Commun. 10, 4952–4971.

Papacharissi, Zizi, 2009. The virtual geographies of social networks: a comparative analysis of Facebook, LinkedIn and ASmallWorld. New Media Soc. 11 (1–2), 199–220.
Robertson, Scott, Vatrapu, Ravi, Medina, Richard, 2010. Off the wall political discourse: Facebook use in the 2008 U.S. presidential election. Inf. Polity 15, 11–31.
Rosenberg, Eli, 2018. Twitter to Tell 677,000 Users they were had by the Russians. Some Signs Show the Problem Continues (Washington Post).
Shah, Neil, Lamba, Hemank, Beutel, Alex, Faloutsos, Christos, 2017. The many faces of link fraud. In: IEEE (Ed.), IEEE International Conference on Data Mining (ICDM). IEEE, pp. 1069–1074.
Stukal, Denis, Sanovich, Sergey, Bonneau, Richard, Tucker, Joshua, 2018. Detecting bots on Russian political twitter. Big Data 5 (4), 310–324.
Tucker, Joshua, Guess, Andrew, Barbera, Pablo, Vaccari, Cristian, Siegel, Alexandra, Sanovich, Sergey, Stukal, Denis, Nyhan, Brendan, 2017. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. William + Flora Hewlett Foundation.
Volkel, J., Feinberg, M., 2017. Morally reframed arguments can affect support for political candidates. Soc. Psychol. Personal. Sci. 9 (8), 917–924.
Yardi, Sarita, Boyd, Danah, 2010. Dynamic debates: an analysis of group polarization over time on twitter. Bull. Sci. Technol. Soc. 30 (5), 316–327.