Exposure to opposing views on social media can increase political polarization

Christopher A. Bail[a,1], Lisa P. Argyle[b], Taylor W. Brown[a], John P. Bumpus[a], Haohan Chen[c], M. B. Fallin Hunzaker[d], Jaemin Lee[a], Marcus Mann[a], Friedolin Merhout[a], and Alexander Volfovsky[e]

[a]Department of Sociology, Duke University, Durham, NC 27708; [b]Department of Political Science, Brigham Young University, Provo, UT 84602; [c]Department of Political Science, Duke University, Durham, NC 27708; [d]Department of Sociology, New York University, New York, NY 10012; and [e]Department of Statistical Science, Duke University, Durham, NC 27708

Edited by Peter S. Bearman, Columbia University, New York, NY, and approved August 9, 2018 (received for review March 20, 2018)

There is mounting concern that social media sites contribute to political polarization by creating "echo chambers" that insulate people from opposing views about current events. We surveyed a large sample of Democrats and Republicans who visit Twitter at least three times each week about a range of social policy issues. One week later, we randomly assigned respondents to a treatment condition in which they were offered financial incentives to follow a Twitter bot for 1 month that exposed them to messages from those with opposing political ideologies (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). Respondents were resurveyed at the end of the month to measure the effect of this treatment, and at regular intervals throughout the study period to monitor treatment compliance. We find that Republicans who followed a liberal Twitter bot became substantially more conservative posttreatment. Democrats exhibited slight increases in liberal attitudes after following a conservative Twitter bot, although these effects are not statistically significant. Notwithstanding important limitations of our study, these findings have significant implications for the interdisciplinary literature on political polarization and the emerging field of computational social science.

political polarization | computational social science | social networks | social media | sociology

Political polarization in the United States has become a central focus of social scientists in recent decades (1–7). Americans are deeply divided on controversial issues such as inequality, gun control, and immigration—and divisions about such issues have become increasingly aligned with partisan identities in recent years (8, 9). Partisan identification now predicts preferences about a range of social policy issues nearly three times as well as any other demographic factor—such as education or age (10). These partisan divisions not only impede compromise in the design and implementation of social policies but also have far-reaching consequences for the effective function of democracy more broadly (11–15).

America's cavernous partisan divides are often attributed to "echo chambers," or patterns of information sharing that reinforce preexisting political beliefs by limiting exposure to opposing political views (16–20). Concern about selective exposure to information and political polarization has increased in the age of social media (16, 21–23). The vast majority of Americans now visit a social media site at least once each day, and a rapidly growing number of them list social media as their primary source of news (24). Despite initial optimism that social media might enable people to consume more heterogeneous sources of information about current events, there is growing concern that such forums exacerbate political polarization because of social network homophily, or the well-documented tendency of people to form social network ties to those who are similar to themselves (25, 26). The endogenous relationship between social network formation and political attitudes also creates formidable challenges for the study of social media echo chambers and political polarization, since it is notoriously difficult to establish whether social media networks shape political opinions, or vice versa (27–29).

Here, we report the results of a large field experiment designed to examine whether disrupting selective exposure to partisan information among Twitter users shapes their political attitudes. Our research is governed by three preregistered hypotheses. The first hypothesis is that disrupting selective exposure to partisan information will decrease political polarization because of intergroup contact effects. A vast literature indicates contact between opposing groups can challenge stereotypes that develop in the absence of positive interactions between them (30). Studies also indicate intergroup contact increases the likelihood of deliberation and political compromise (31–33). However, all of these previous studies examine interpersonal contact between members of rival groups. In contrast, our experiment creates virtual contact between members of the public and opinion leaders from the opposing political party on a social media site. It is not yet known whether such virtual contact creates the

Significance

Social media sites are often blamed for exacerbating political polarization by creating "echo chambers" that prevent people from being exposed to information that contradicts their preexisting beliefs. We conducted a field experiment that offered a large group of Democrats and Republicans financial compensation to follow bots that retweeted messages by elected officials and opinion leaders with opposing political views. Republican participants expressed substantially more conservative views after following a liberal Twitter bot, whereas Democrats' attitudes became slightly more liberal after following a conservative Twitter bot—although this effect was not statistically significant. Despite several limitations, this study has important implications for the emerging field of computational social science and ongoing efforts to reduce political polarization online.

Author contributions: C.A.B., L.P.A., T.W.B., J.P.B., H.C., M.B.F.H., J.L., M.M., F.M., and A.V. designed research; C.A.B., L.P.A., T.W.B., H.C., M.B.F.H., J.L., M.M., and F.M. performed research; C.A.B., T.W.B., H.C., J.L., and A.V. contributed new reagents/analytic tools; C.A.B., L.P.A., T.W.B., H.C., M.B.F.H., J.L., M.M., F.M., and A.V. analyzed data; and C.A.B., L.P.A., T.W.B., M.B.F.H., M.M., F.M., and A.V. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All data, code, and the markdown file used to create this report will be available at this link on the Dataverse: https://dataverse.harvard.edu/dataverse.xhtml?alias=chris_bail.

[1]To whom correspondence should be addressed. Email: christopher.bail@duke.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1804840115/-/DCSupplemental.

Published online August 28, 2018.

same type of positive mutual understanding—or whether the relative anonymity of social media forums emboldens people to act in an uncivil manner. Such incivility could be particularly rife in the absence of facial cues and other nonverbal gestures that might prevent the escalation of arguments in offline settings.

Our second hypothesis builds upon a more recent wave of studies that suggest exposure to those with opposing political views may create backfire effects that exacerbate political polarization (34–37). This literature—which now spans several academic disciplines—indicates people who are exposed to messages that conflict with their own attitudes are prone to counterargue them using motivated reasoning, which accentuates perceived differences between groups and increases their commitment to preexisting beliefs (34–37). Many studies in this literature observe backfire effects via survey experiments where respondents are exposed to information that corrects factual inaccuracies—such as the notion that Saddam Hussein possessed weapons of mass destruction prior to the 2003 US invasion of Iraq—although these findings have failed to replicate in two recent studies (38, 39). Yet our study is not designed to evaluate attempts to correct factual inaccuracies. Instead, we aim to assess the broader impact of prolonged exposure to counterattitudinal messages on social media.

Our third preregistered hypothesis is that backfire effects will be more likely to occur among conservatives than liberals. This hypothesis builds upon recent studies that indicate conservatives hold values that prioritize certainty and tradition, whereas liberals value change and diversity (40, 41). We also build upon recent studies in cultural sociology that examine the deeper cultural schemas and narratives that create and sustain such value differences (34, 26). Finally, we also build upon studies that observe asymmetric polarization in roll call voting wherein Republicans have become substantially more conservative whereas Democrats exhibit little or no increase in liberal voting positions (42). Although a number of studies have found evidence of this trend, we are not aware of any that examine such dynamics among the broader public—and on social media in particular.

Research Design

Fig. 1 provides an overview of our research design. We hired a professional survey firm to recruit self-identified Republicans and Democrats who visit Twitter at least three times each week to complete a 10-min survey in mid-October 2017 and 1.5 mo later. These surveys measure the key outcome variable: change in political ideology during the study period via a 10-item survey instrument that asked respondents to agree or disagree with a range of statements about policy issues on a seven-point scale ($\alpha = .91$) (10). Our survey also collected information about other political attitudes, use of social media and conventional media sources, and a range of demographic indicators that we describe in *SI Appendix*. Finally, all respondents were asked to report their Twitter ID, which we used to mine additional information about their online behavior, including the partisan background of the accounts they follow on Twitter. Our research was approved by the Institutional Review Boards at Duke University and New York University. All respondents provided informed consent before participating in our research.

We ran separate field experiments for Democratic and Republican respondents, and, within each group, we used a block randomization design that further stratified respondents according to two variables that have been linked to political polarization: (*i*) level of attachment to political party and (*ii*) level of interest in current events. We also randomized assignment according to respondents' frequency of Twitter use, which we reasoned would influence the amount of exposure to the intervention we describe in the following paragraph and thereby the overall likelihood of opinion change.

We received 1,652 responses to our pretreatment survey (901 Democrats and 751 Republicans). One week later, we randomly assigned respondents to a treatment condition, thus using an "ostensibly unrelated" survey design (43). At this time, respondents in the treatment condition were offered $11 to follow a Twitter bot, or automated Twitter account, that they were told would retweet 24 messages each day for 1 mo. Respondents were not informed of the content of the messages the bots would retweet. As Fig. 2 illustrates, we created a liberal Twitter bot and a conservative Twitter bot for each of our experiments. These bots retweeted messages randomly sampled from a list of 4,176 political Twitter accounts (e.g., elected officials, opinion leaders, media organizations, and nonprofit groups). These accounts were identified via a network-sampling technique that assumes those with similar political ideologies are more likely to follow each other on Twitter than those with opposing political ideologies (44). For further details about the design of the study's bots, please refer to *SI Appendix*.

To monitor treatment compliance, respondents were offered additional financial incentives (up to $18) to complete weekly surveys that asked them to answer questions about the content of the tweets produced by the Twitter bots and identify a picture of an animal that was tweeted twice a day by the bot but deleted immediately before the weekly survey. At the conclusion of the study period, respondents were asked to complete a final survey with the same questions from the initial (pretreatment) survey. Of those invited to follow a Twitter bot, 64.9% of Democrats and 57.2% of Republicans accepted our invitation. Approximately 62% of Democrats and Republicans who followed the bots were able to answer all substantive questions about the content of messages retweeted each week, and 50.2% were able to identify the animal picture retweeted each day.

Results

Fig. 3 reports the effect of being assigned to the treatment condition, or the Intent-to-Treat (ITT) effects, as well as the Complier Average Causal Effects (CACE) which account for the differential rates of compliance among respondents we observed. These estimates were produced via multivariate models that predict respondents' posttreatment scores on the liberal/conservative scale described above, controlling for pretreatment scores on this scale as well as 12 other covariates described in *SI Appendix*. We control for respondents' pretreatment liberal/conservative scale score to mitigate the influence of period effects. Negative scores indicate respondents became more liberal in response to treatment, and positive scores indicate they became more conservative. Circles describe unstandardized point estimates, and the horizontal lines in Fig. 3 describe 90% and 95% confidence intervals. We measured compliance with treatment in three ways. "Minimally Compliant Respondents" describes those who followed our bot throughout the entire study period. "Partially Compliant Respondents" are those who were able to answer at least one—but not all—questions about the content of one of the bots' tweets administered each week during the survey period. "Fully Compliant Respondents" are those who successfully answered all of these questions. These last two categories are mutually exclusive.

Although treated Democrats exhibited slightly more liberal attitudes posttreatment that increase in size with level of compliance, none of these effects were statistically significant. Treated Republicans, by contrast, exhibited substantially more conservative views posttreatment. These effects also increase with level of compliance, but they are highly significant. Our most cautious estimate is that treated Republicans increased 0.12 points on a seven-point scale, although our model that estimates the effect of treatment upon fully compliant respondents indicates this effect

Initial Survey

Respondents were offered $11 to provide their Twitter ID and complete a 10-minute survey about their political attitudes, social media use, and media consumption habits (demographics provided by survey firm).

Randomization

One week later, respondents were assigned to treatment and control conditions within strata created using pre-treatment covariates that describe attachment to party, frequency of Twitter use, and overall interest in current events.

Weekly Surveys

Respondents in treatment conditions informed they are eligible to receive up to $6 each week during the study period for correctly answering questions about the content of messages retweeted by Twitter .Bots.

Post-Survey

Respondents were offered $12 to repeat the pre-treatment survey one month after initial survey.

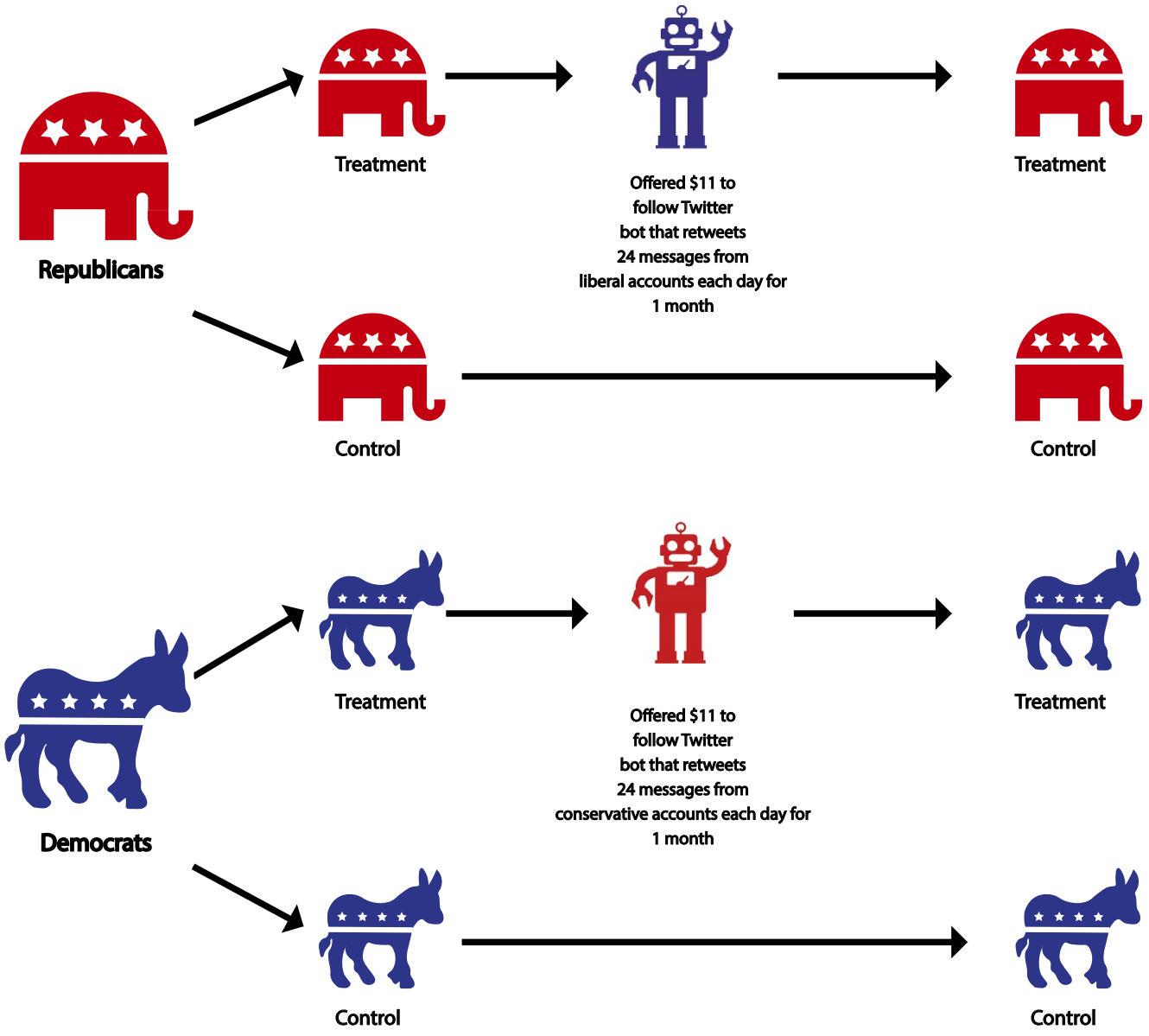


Fig. 1. Overview of research design.

is substantially larger (0.60 points). These estimates correspond to an increase in conservatism between 0.11 and 0.59 standard deviations.

Discussion and Conclusion

Before discussing the implications of these findings, we first note important limitations of our study. Readers should not interpret

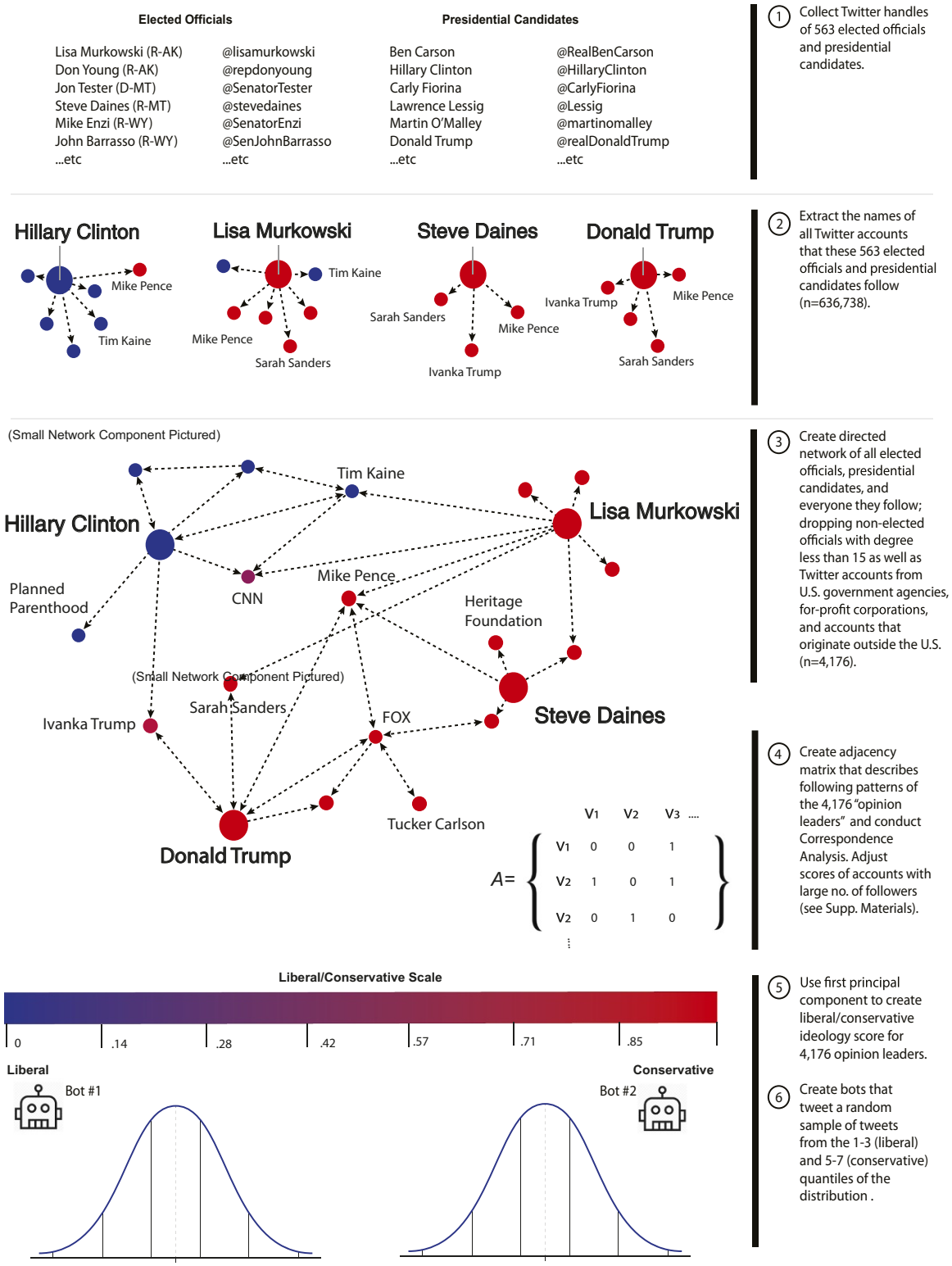


Fig. 2. Design of study's Twitter bots.

our findings as evidence that exposure to opposing political views will increase polarization in all settings. Although ours is among the largest field experiments conducted on social media to date, the findings above should not be generalized to the entire US population, because a majority of Americans do not use Twitter (24). It is also unclear how exposure to opposing views might shape political polarization in other parts of the world. In addition, we did not study people who identify as independents, or those who use Twitter but do so infrequently. Such individuals might exhibit quite different reactions to an intervention such as ours. Future studies are needed to further evaluate the external validity of our findings, because we offered our respondents

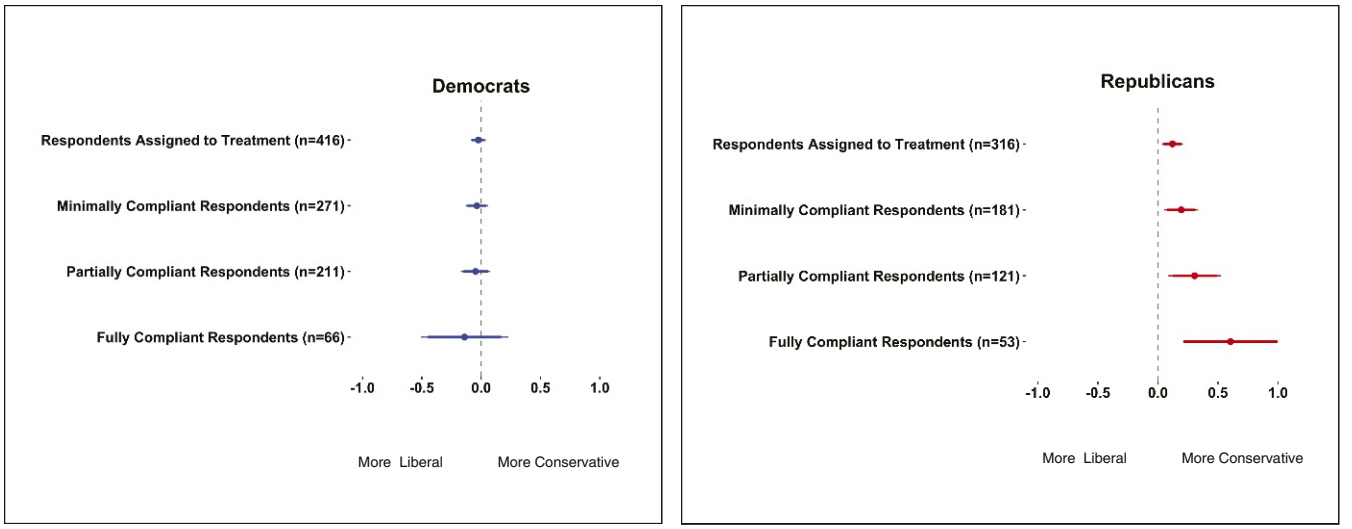


Fig. 3. Effect of following Twitter bots that retweet messages by elected officials, organizations, and opinion leaders with opposing political ideologies for 1 mo, on a seven-point liberal/conservative scale where larger values indicate more conservative opinions about social policy issues, for experiments with Democrats (n = 697) and Republicans (n = 542). Models predict posttreatment liberal/conservative scale score and control for pretreatment score on this scale as well as 12 other covariates described in *SI Appendix*. Circles describe unstandardized point estimates, and bars describe 90% and 95% confidence intervals. "Respondents Assigned to Treatment" describes the ITT effect for Democrats (ITT = −0.02, t = −0.76, p = 0.45, n = 416) and Republicans (ITT = 0.12, t = 2.68, p = 0.008, n = 316). "Minimally-Compliant Respondents" describes the CACE for respondents who followed one of the study's bots for Democrats (CACE = −0.04, t = −0.75, p = 0.45, n of compliant respondents = 271) and Republicans (CACE = 0.19, t = 2.73, $p < 0.007$, n of compliant respondents = 181). "Partially-Compliant Respondents" describes the CACE for respondents who correctly answered at least one question, but not all questions, about the content of a bot's tweets during weekly surveys throughout the study period for Democrats (CACE = −0.05, t = −0.75, p = 0.45, n of compliant respondents = 211) and Republicans (CACE = 0.31, t = 2.73, $p <.007$, n of compliant respondents = 121). "Fully-Compliant Respondents" describes the CACE for respondents who answered all questions about the content of the bot's tweets correctly for Democrats (CACE = −0.14, t = −0.75, p = 0.46, n of compliant respondents = 66) and Republicans (CACE = 0.60, t = 2.53, $p < 0.01$, n of compliant respondents = 53). Although treated Democrats exhibited slightly more liberal attitudes posttreatment that increase in size with level of compliance, none of these effects were statistically significant. In contrast, treated Republicans exhibited substantially more conservative views posttreatment that increase in size with level of compliance, and these effects are highly significant.

financial incentives to read messages from people or organizations with opposing views. It is possible that Twitter users may simply ignore such counterattitudinal messages in the absence of such incentives. Perhaps the most important limitation of our study is that we were unable to identify the precise mechanism that created the backfire effect among Republican respondents reported above. Future studies are thus urgently needed not only to determine whether our findings replicate in different populations or within varied social settings but to further identify the precise causal pathways that create backfire effects more broadly.

Future studies are also needed because we cannot rule out all alternative explanations of our findings. In *SI Appendix*, we present additional analyses that give us confidence that our results are not driven by Hawthorne effects, partisan "learning" processes, variation in the ideological extremity of messages by party, or demographic differences in social media use by age. At the same time, we are unable to rule out other alternative explanations discussed in *SI Appendix*. For example, it is possible that our findings resulted from increased exposure to information about politics, and not exposure to opposing messages per se. Similarly, increases in conservatism among Republicans may have resulted from increased exposure to women or racial and ethnic minorities whose messages were retweeted by our liberal bot. Finally, our intervention only exposed respondents to high-profile elites with opposing political ideologies. Although our liberal and conservative bots randomly selected messages from across the liberal and conservative spectrum, previous studies indicate such elites are significantly more polarized than the general electorate (45). It is thus possible that the backfire effect we identified could be exacerbated by an antielite bias, and future studies are needed to examine the effect of online intergroup contact with nonelites.

Despite these limitations, our findings have important implications for current debates in sociology, political science, social psychology, communications, and information science. Although we found no evidence that exposing Twitter users to opposing views reduces political polarization, our study revealed significant partisan differences in backfire effects. This finding is important, since our study examines such effects in an experimental setting that involves repeated contact between rival groups across an extended time period on social media. Our field experiment also disrupts selective exposure to information about politics in a real-world setting through a combination of survey research, bot technology, and digital trace data collection. This methodological innovation enabled us to collect information about the nexus of social media and politics with high granularity while developing techniques for measuring treatment compliance, mitigating causal interference, and verifying survey responses with behavioral data—as we discuss in *SI Appendix*. Together, we believe these contributions represent an important advance for the nascent field of computational social science (46).

Although our findings should not be generalized beyond party-identified Americans who use Twitter frequently, we note that recent studies indicate this population has an outsized influence on the trajectory of public discussion—particularly as the media itself has come to rely upon Twitter as a source of news and a window into public opinion (47). Although limited in scope, our findings may be of interest to those who are working to reduce political polarization in applied settings. More specifically, our study indicates that attempts to introduce people to a broad range of opposing political views on a social media site such as Twitter might be not only be ineffective but counterproductive—particularly if such interventions are initiated by liberals. Since previous studies have produced substantial evidence that intergroup contact produces compromise and mutual understanding in other contexts, however, future attempts to reduce political

polarization on social media will most likely require learning which types of messages, tactics, or issue positions are most likely to create backfire effects and whether others—perhaps delivered by nonelites or in offline settings—might be more effective vehicles to bridge America's partisan divides.

Materials and Methods

See *SI Appendix* for a detailed description of all materials and methods used within this study as well as links to our preregistration statement, replication materials, additional robustness checks, and an extended discussion of alternative explanations of our findings. Our research was approved by the Institutional Review Boards at Duke University and New York University.

ACKNOWLEDGMENTS. We thank Paul DiMaggio, Sunshine Hillygus, Gary King, Fan Li, Arthur Lupia, Brendan Nyhan, and Samantha Luks for helpful conversations about this study prior to our research. Our work was supported by the Carnegie Foundation, the Russell Sage Foundation, and the National Science Foundation.

1. DiMaggio P, Evans J, Bryson B (1996) Have American's social attitudes become more polarized? *Am J Sociol* 102:690–755.
2. Iyengar S, Westwood SJ (2015) Fear and loathing across party lines: New evidence on group polarization. *Am J Polit Sci* 59:690–707.
3. Baldassarri D, Gelman A (2008) Partisans without constraint: Political polarization and trends in American public opinion. *Am J Sociol* 114:408–446.
4. Sides J, Hopkins DJ (2015) *Political Polarization in American Politics* (Bloomsbury, London).
5. Baldassarri D, Bearman P (2007) Dynamics of political polarization. *Am Sociol Rev* 72:784–811.
6. Fiorina MP, Abrams SJ (2008) Political polarization in the American public. *Annu Rev Polit Sci* 11:563–588.
7. DellaPosta D, Shi Y, Macy M (2015) Why do liberals drink lattes? *Am J Sociol* 120: 1473–1511.
8. Levendusky M (2009) *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans* (Univ Chicago Press, Chicago).
9. Mason L (2018) *Uncivil Agreement: How Politics Became Our Identity* (Univ Chicago Press, Chicago).
10. Dimock M, Carroll D (2014) Political polarization in the American public: How increasing ideological uniformity and partisan antipathy affect politics, compromise, and everyday life (Pew Res Cent, Washington, DC).
11. Achen CH, Bartels LM (2016) *Democracy for Realists: Why Elections Do Not Produce Responsive Government* (Princeton Univ Press, Princeton).
12. Erikson RS, Wright GC, McIver JP (1993) *Public Opinion and Policy in the American States* (Cambridge Univ Press, Cambridge, UK).
13. Fishkin JS (2011) *When the People Speak: Deliberative Democracy and Public Consultation* (Oxford Univ Press, Oxford).
14. Bennett WL, Iyengar S (2008) A new era of minimal effects? The changing foundations of political communication. *J Commun* 58:707–731.
15. Sunstein C (2002) *Republic.com* (Princeton Univ Press, Princeton).
16. Bakshy E, Messing S, Adamic LA (2015) Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348:1130–1132.
17. Sunstein C (2001) *Echo Chambers, Bush v. Gore, Impeachment, and Beyond* (Princeton Univ Press, Princeton).
18. King G, Schneer B, White A (2017) How the news media activate public expression and influence national agendas. *Science* 358:776–780.
19. Berry JM, Sobieraj S (2013) *The Outrage Industry: Political Opinion Media and the New Incivility* (Oxford Univ Press, Oxford).
20. Prior M (2013) Media and political polarization. *Annu Rev Polit Sci* 16: 101–127.
21. Pariser E (2011) *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (Penguin, New York).
22. Conover M, Ratkiewicz J, Francisco M (2011) Political polarization on twitter. *ICWSM* 133:89–96.
23. Boxell L, Gentzkow M, Shapiro JM (2017) Greater Internet use is not associated with faster growth in political polarization among us demographic groups. *Proc Natl Acad Sci USA* 114:10612–10617.
24. Perrin A (2015) Social media usage: 2005-2015 (Pew Res Cent, Washington, DC).
25. McPherson M, Smith-lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annu Rev Sociol* 27:415–444.
26. Edelmann A, Vaisey S (2014) Cultural resources and cultural distinction in networks. *Poetics* 46:22–37.
27. Lazer D, Rubineau B, Chetkovich C, Katz N, Neblo M (2010) The coevolution of networks and political attitudes. *Polit Commun* 27:248–274.
28. Centola D (2011) An experimental study of homophily in the adoption of health behavior. *Science* 334:1269–1272.
29. Vaisey S, Lizardo O (2010) Can cultural worldviews influence network composition?. *Social Forces* 88:1595–1618.
30. Pettigrew TF, Tropp LR (2006) A meta-analytic test of intergroup contact theory. *J Pers Soc Psychol* 90:751–783.
31. Huckfeldt R, Johnson PE, Sprague J (2004) *Political Disagreement: The Survival of Diverse Opinions Within Communication Networks* (Cambridge Univ Press, Cambridge, UK).
32. Mutz DC (2002) Cross-cutting social networks: Testing democratic theory in practice. *Am Polit Sci Rev* 96:111–126.
33. Grönlund K, Herne K, Setälä M (2015) Does enclave deliberation polarize opinions? *Polit Behav* 37:995–1020.
34. Bail C (2015) *Terrified: How Anti-Muslim Fringe Organizations Became Mainstream* (Princeton Univ Press, Princeton).
35. Lord CG, Ross L, Lepper MR (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J Pers Soc Psychol* 37:2098–2109.
36. Nyhan B, Reifler J (2010) When corrections fail: The persistence of political misperceptions. *Polit Behav* 32:303–330.
37. Taber CS, Lodge M (2006) Motivated skepticism in the evaluation of political beliefs. *Am J Polit Sci* 50:755–769.
38. Wood T, Porter E (2016) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Polit Behav*, 1–29.
39. Wood T, Porter E (2018) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Polit Behav*, 10.1007/s11109-018-9443-y.
40. Graham J, Haidt J, Nosek BA (2009) Liberals and conservatives rely on different sets of moral foundations. *J Pers Social Psychol* 96:1029–1046.
41. Jost JT, et al. (2007) Are needs to manage uncertainty and threat associated with political conservatism or ideological extremity? *Pers Soc Psychol Bull* 33:989–1007.
42. Grossmann M, Hopkins DA (2016) *Asymmetric Politics: Ideological Republicans and Group Interest Democrats* (Oxford Univ Press, Oxford).
43. Broockman D, Kalla J (2016) Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* 352:220–224.
44. Barberá P (2014) Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Polit Anal* 23:76–91.
45. Abramowitz AI, Saunders KL (2008) Is polarization a myth?. *J Polit* 70:542–555.
46. Lazer D, et al (2009) Life in the network: The coming age of computational social science. *Science* 323:721–723.
47. Faris R, et al. (2017) Partisanship, Propaganda, and disinformation: Online media and the 2016 U.S. presidential election, (Berkman Klein Cent Internet Soc Harvard Univ, Cambridge, MA).

Supplementary Materials for "Exposure to Opposing Views can Increase Political Polarization: Evidence from a Large-scale Field Experiment on Social Media"

Christopher A. Bail, Lisa Argyle, Taylor W. Brown, John Bumpus, Haohan Chen, M.B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, Alexander Volfovsky

07/03/18

Contents

www.pnas.org/cgi/doi/10.1073/pnas.1804840115

1 Introduction

1.1 Replication Materials

This document describes all materials and methods for the article "Exposure to Opposing Views can Increase Political Polarization: Evidence from a Large-scale Field Experiment on Social Media," by Bail et al. All data, code, and the markdown file used to create this report will be available at this link on the Dataverse.

1.2 Pre-registration

The research design and hypotheses described in the main text of our article were pre-registered via the Open Science Framework and can be accessed at this link.

2 Pre-Treatment Survey and Randomization

2.1 Power Analyses

In order to identify a suitable sample size to evaluate our hypotheses, we conducted a literature review for studies about the relationship between exposure to political outgroups and political polarization. The study most similar to our own that we were able to identify was Grönlund et al's (2015) article, "Does Enclave Deliberation Polarize Opinions?," which appeared in the journal *Political Behavior*. This study presents a field experiment to gauge opinions about immigration in Finland. Out of a total sample of 805 respondents, 366 were assigned to a treatment condition in which they were either asked to deliberate about immigration with a group of people who had heterogeneous views about immigration or a control group condition where they were assinged to groups whose members held views that were little or no different than their own. Grönlund et al. report the attitudes of those in the four treatment conditions showed an average increase in positive attitudes towards immigrants between 0.29 and 1.8 on a scale of 0 to 14 (SD: 2.98)—equal to 0.1 and 0.6 standard deviations.

Figure 1 below reports a power analysis for our study if the effect size were identical to the largest effect reported by Grönlund et al. The red line describes the 80% criterion for probability of sufficient statistical power that is widely employed across the social sciences. According to this calculation, our study would only require approximately 100 people for treatment and control samples. Because our study has two treatment conditions, this would indicate a sample size of 400 is warranted. Yet there are two important differences between this study and our own. First, we planned to expose respondents to out-group Twitter messages for one month, whereas Grönlund et al.'s (2015) study occurred over a single weekend. Second, our study involves virtual contact between respondents and out-groups on a social media site, whereas Grönlund's study involved in-person deliberation.

Though one could argue that the length of our treatment balances out the greater intimacy of Grönlund et al.'s intervention, we believe a more conservative approach is warranted because ours is the first study to examine this process on social media. Compared to in-person deliberation, previous studies indicate online

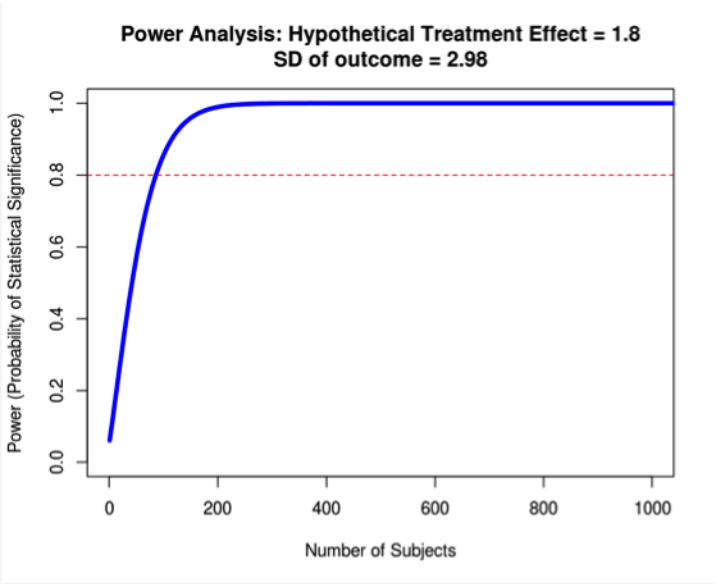


Figure 1: Power Analysis #1.

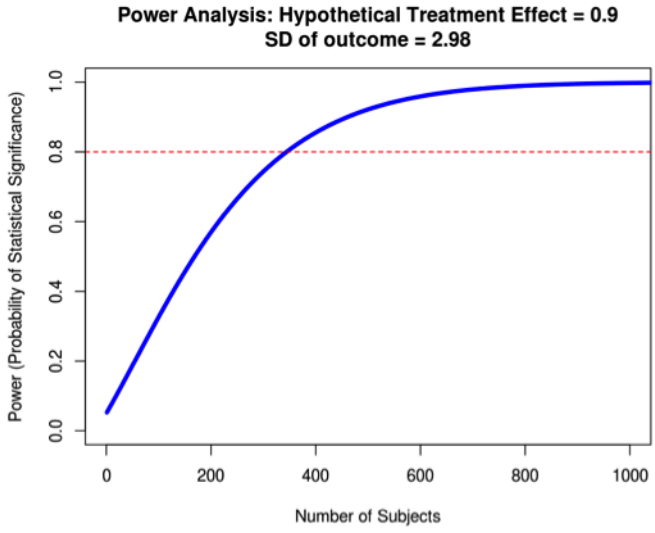


Figure 2: Power analysis #2.

interventions tends to produce smaller attitudinal changes (Luskin, Fishkin, and Hahn 2007). Therefore, we performed an additional power analysis (pictured in Figure 2 below) that estimated ideal sample sizes if the effect we observe is half the size of the largest effect from Grönlund et al.'s (2015) study. This illustration indicated that approximately 300 respondents are required per treatment/control comparison (or 1,200 respondents overall between the Democratic and Republican experiments).

2.2 Survey Recruitment Process

We hired YouGov—one of the largest and most reputable online survey firms in the United States—to recruit at least 1,200 self-identified Republicans and Democrats over age 17 who visit Twitter at least three times each week to complete five surveys between mid-October 2017 and mid-November 2017. A more detailed description of YouGov and its web panelists is available here. Figure 3 provides a detailed description of the recruitment process for our pre-treatment survey, which was fielded between October 10th and October 19th, 2017. YouGov invited 10,634 members of its U.S. panel to participate in our study using U.S. census sampling frames. Of these, 5,520 did not respond, and 5,114 accepted the invitation, for an initial cooperation rate of 48% (AAPOR RR3 = 42.7%). These individuals were then asked several screening questions. First, they were asked about their party identification using the following question: "Generally speaking, do you think of yourself as a [Democrat/Republican/Independent/Other/Not Sure]." Respondents who did not respond with either "Democrat" or "Republican" were screened out, and remaining respondents were asked the following question, which was used to identify the treatment blocks described in further detail below: "Would you call yourself a strong Democrat/Republican or a not very strong Democrat/Republican?" These two questions have been widely employed to measure party attachment in the American National Election Study and many other surveys. Third, respondents were asked if they "visit Twitter at least three times a week in order to read messages from other Twitter accounts," and screened out if they answered negatively.

A total of 2,539 people were deemed eligible according to these two initial eligibility criteria, and were subsequently re-directed to an informed consent dialogue and offered the equivalent of $11 via YouGov's "points" system, which allows respondents to redeem points for items such as Amazon gift cards, to share their Twitter handle, or Twitter ID, in order that it may be linked to their survey responses.

In the informed consent dialogue pictured in Figure 4, eligible pre-treatment survey participants were informed that our survey would take about 10 minutes and was designed to "investigate peoples' experiences on Twitter." They were also informed that participants who completed this survey and provided a valid Twitter handle would be eligible to complete a follow-up survey one month later.

1,754 respondents agreed and completed the entire pre-treatment survey. 500 respondents began—but did not complete—the pre-treatment survey, and 285 respondents refused to complete the pre-treatment survey. Of the 1,754 respondents that completed the pre-treatment survey, 102 were removed by YouGov's quality algorithm, which eliminates respondents who complete the survey within a time frame that is deemed impossible by the algorithm. This resulted in an initial sample of 1,652 respondents.

One month later, respondents who were not eliminated prior to treatment assignment for one of the reasons described in the next section were invited to complete the post-treatment survey. In the informed consent for the post-treatment survey, participants were reminded "One month ago, you completed a survey to investigate your experiences on Twitter...", and were informed that they were now invited to complete a 10 minute follow up survey about "what you think about important issues and how you use Twitter and other media sources."

2.3 Respondents Eliminated Before Treatment Assignment

136 of the 1,652 respondents who completed the pre-treatment survey were excluded from subsequent analyses because they did not present a valid Twitter handle or username that could be accessed via Twitter's Application Programming Interface. Forty-five respondents were excluded because they provided poor quality data, indicated by providing the same answer to ten consecutive questions that were randomized according

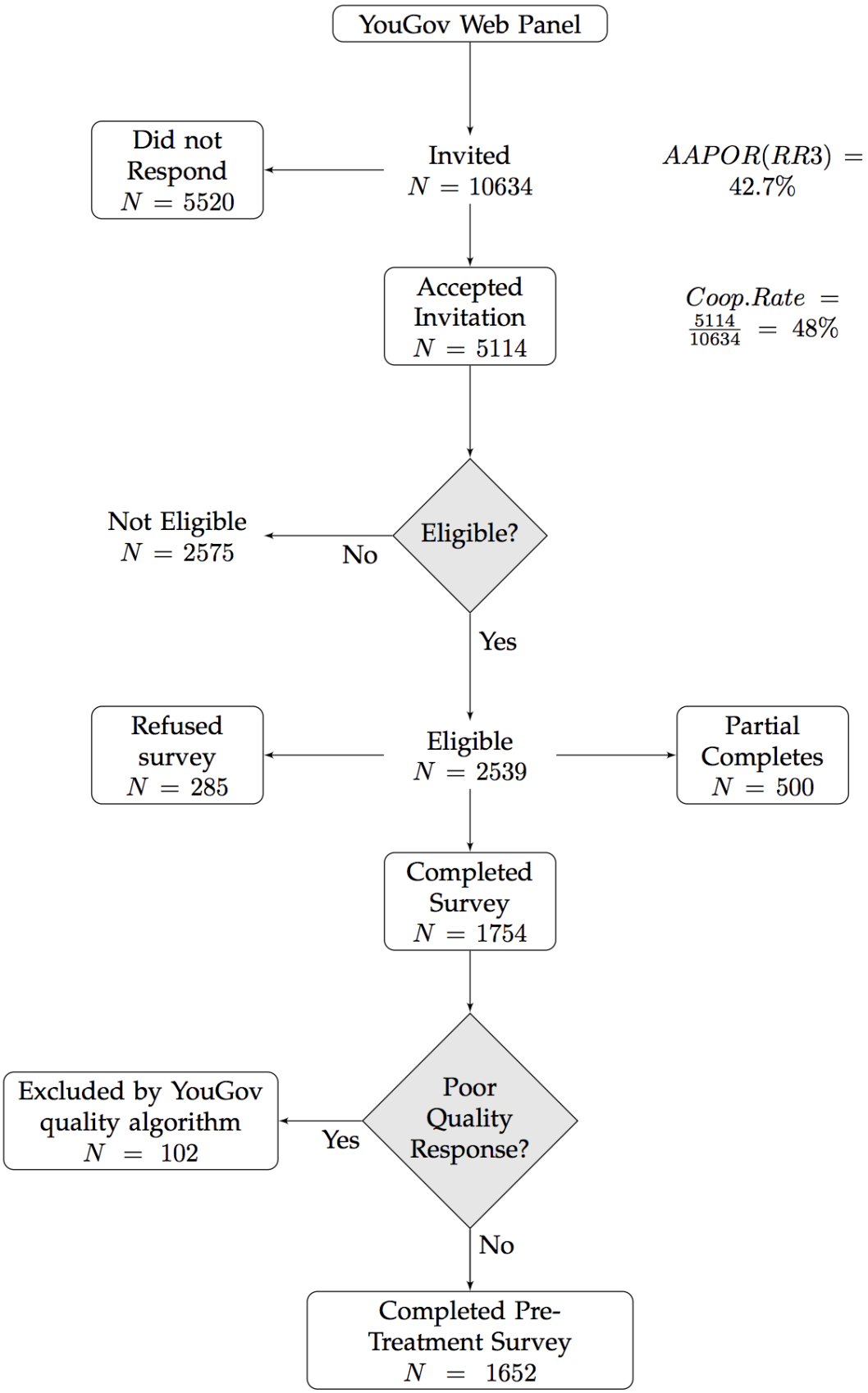


Figure 3: Recruitment Process for Pre-Treatment Survey

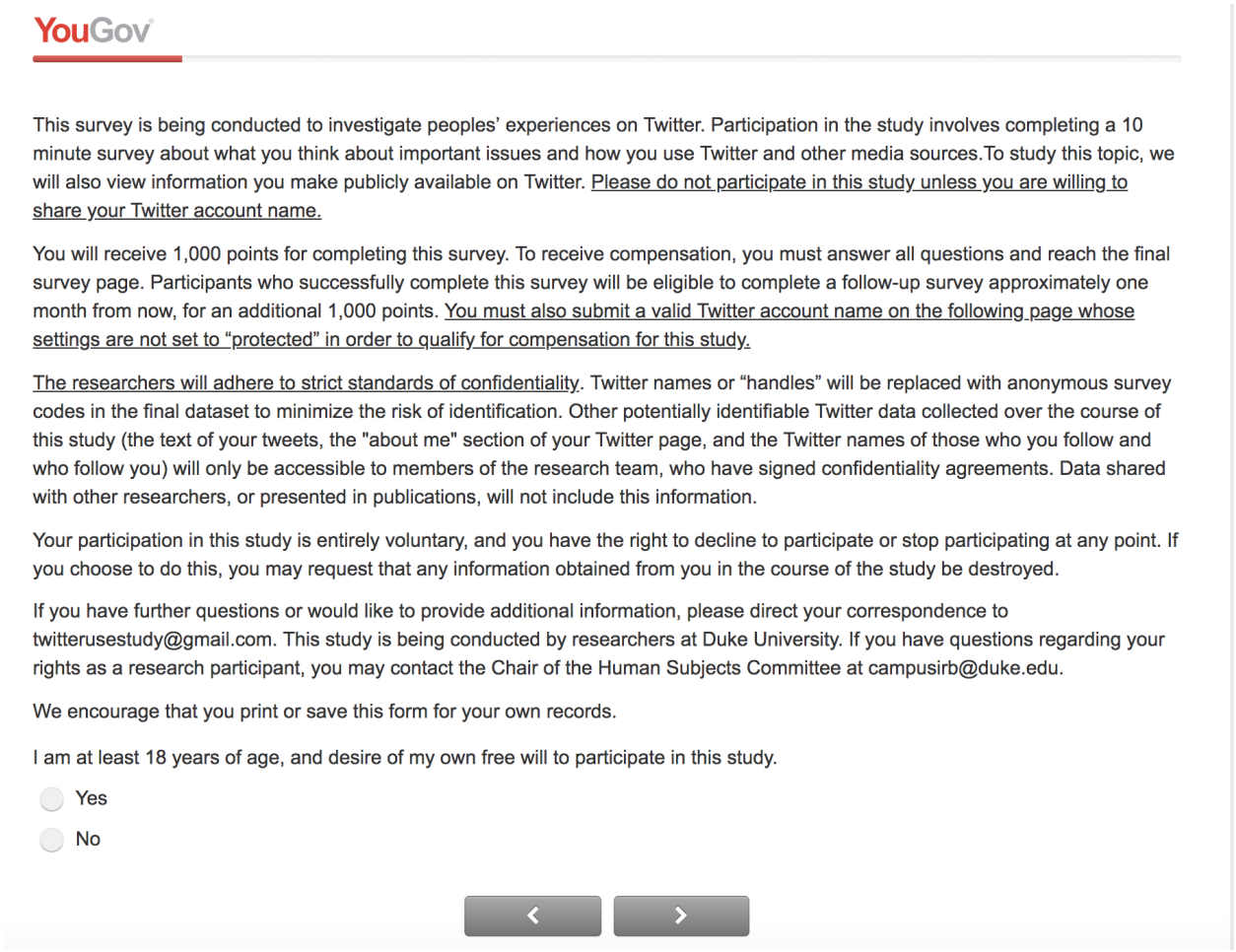


Figure 4: Informed Consent Dialogue

to whether the respondent was asked to agree with a liberal or conservative-leaning statement (an additional twelve respondents were later excluded because they did the same during the final post-treatment survey).

A significant advantage of our research design is that we were able to cross-validate survey responses with behavioral and demographic information available from the Twitter profiles and messages of respondents to our pre-treatment survey. Forty-four respondents were excluded because they did not follow any accounts on Twitter, and therefore did not satisfy our screening criterion that participants be active Twitter users who "regularly log on to read messages from other Twitter accounts." We elected to exclude any respondent for whom demographic information in the survey conflicted with at least two demographic variables that were observable on the respondent's Twitter page (age, gender, race, and geographic location). Some of these respondents were excluded because of the aforementioned exclusion rules, but an additional 74 respondents were dropped because they provided highly inconsistent information in the survey and in their Twitter profile. 4 additional respondents were excluded because we suspected they provided an account of a famous person instead of their own Twitter account. We operationalized fame as having more than 50,000 followers. Because it was theoretically possible that a respondent in our study could have a large number of followers, we cross-referenced demographic information from the Twitter account in question with that reported in our survey and identified significant discrepancies which further increased our confidence that these responses were non-valid.

2.4 Causal Interference

Yet another advantage of our research design is that we were able to collect social network data from each respondent's Twitter account in order to mitigate the risk of causal interference in our survey population. For example, respondents in our control condition could receive partial treatment if they follow a respondent in the treatment condition who retweeted—or otherwise engaged with—a message produced by one of the bots created by our research team. After removing respondents who were excluded for reasons described in the previous section of this document, we identified 136 respondents in our sample who followed—or were followed by—at least one other respondent in our study. As Figure 5 shows, 90 of these people were part of network components that included at least two other participants in the study. We excluded all respondents that were part of such components from the current study before treatment assignment, but treated some of them as part of a separate study designed to gauge opinion-leader dynamics that we will report at a later date. Of the remaining 46 people who were connected to only one other person in the survey population, we randomly dropped one respondent within each dyad before treatment assignment.

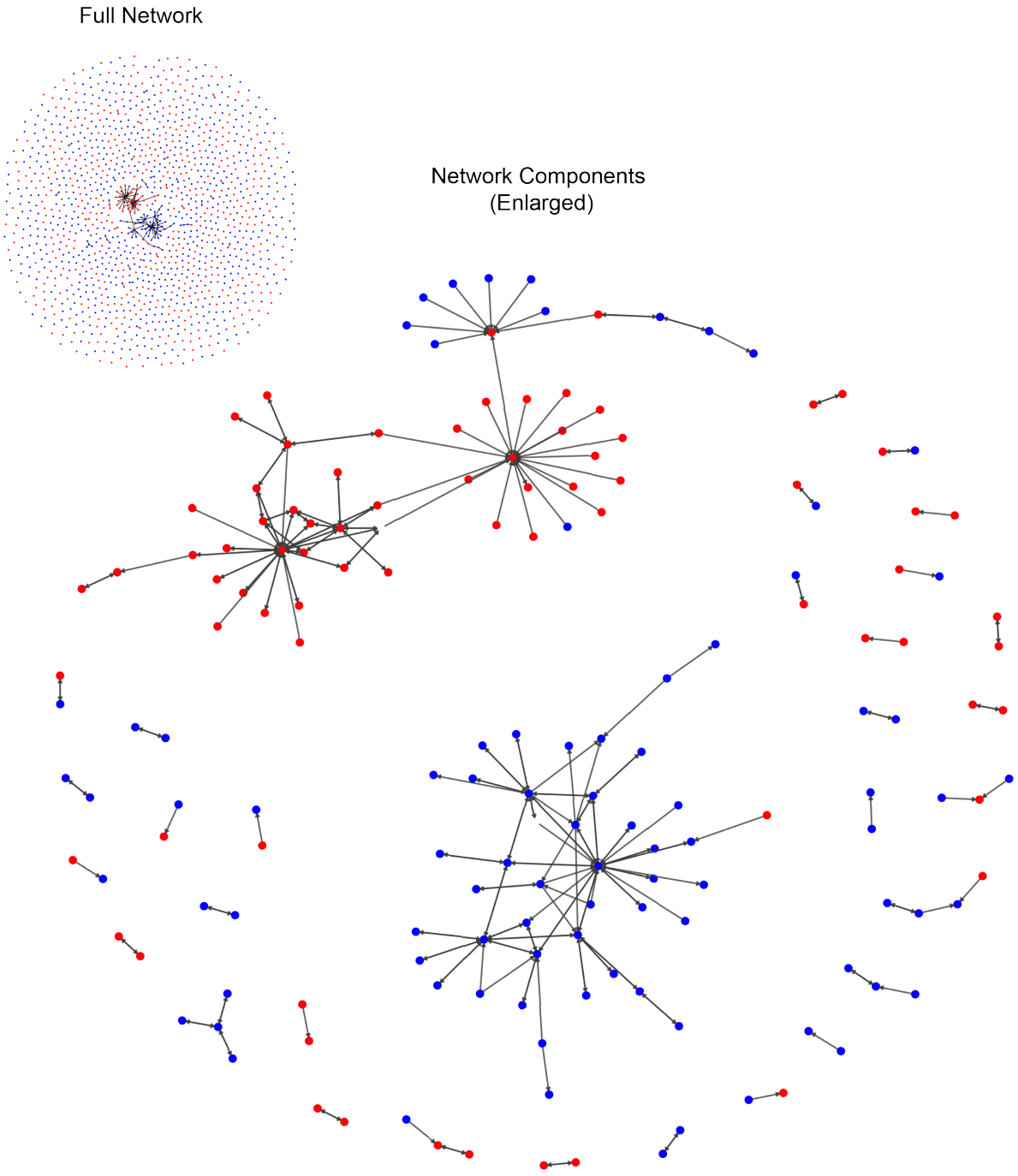


Figure 5: Network Diagram Describing Twitter Connections between Respondents in Initial Pre-Treatment Survey

2.5 Descriptive Characteristics of Final Study Population

Our final sample included 1,220 respondents (691 self-identified Democrats, and 529 self-identified Republicans). Table 1 below compares our sample to data from the 2016 American Community Survey, which are available here. Data on state populations for 2016 were collected from the U.S. Census, and are available here. As this Table shows, our sample closely approximates the adult population distribution across geographic regions of the United States, races and ethnicities, and gender. Table 1 also provides descriptive characteristics of both Democratic and Republican respondents in our study.

Table 1: Comparison of Demographic Characteristics of Respondents in Study Sample to U.S. Census/American Community Survey

Variable	National Mean	Study Mean	p	Study Dem. Mean	Study Rep. Mean	p
Age	37.84	50.49	0.87	50.31	50.72	0.64
Female	0.51	0.52	0.48	0.55	0.48	0.02
White	0.70	0.84	0.00	0.78	0.92	0.00
Asian	0.05	0.01	0.00	0.01	0.01	0.38
Black	0.12	0.06	0.00	0.10	0.02	0.00
Native American	0.01	0.01	0.89	0.01	0.01	0.77
Other	0.05	0.05	0.83	0.06	0.04	0.04
Mixed	0.06	0.02	0.00	0.03	0.01	0.01
New England	0.03	0.03	1.00	0.04	0.03	0.55
Mid-Atlantic	0.13	0.15	0.00	0.17	0.13	0.05
East North Central	0.14	0.14	0.87	0.15	0.13	0.62
West North Central	0.06	0.06	0.43	0.06	0.05	0.55
South Atlantic	0.20	0.22	0.05	0.20	0.24	0.15
East South Central	0.06	0.07	0.24	0.06	0.08	0.29
West South Central	0.12	0.10	0.01	0.08	0.12	0.01
Mountain	0.07	0.06	0.02	0.05	0.06	0.56
Pacific	0.18	0.17	0.40	0.19	0.15	0.11

2.6 Block Randomization

In order to ensure sufficient statistical power, we employed a block randomized design. We identified two strong covariates of political polarization based upon a comprehensive review of the literature. These included a) level of attachment to political party (measured using a binary "strong/not very strong Republican/Democrat" indicator employed by the American National Election Study); and, b) overall level of interest in news and current events (those who described themselves as interested "most of the time"). We also stratified by frequency of Twitter usage, because we reasoned that more frequent Twitter users would be exposed to more of the messages produced by our study's Twitter bots. Our cutoff binned respondents according to whether they visit Twitter at least one time per day or not. One week after the pre-treatment survey, all respondents who provided valid responses were randomized to treatment using the aforementioned design.

To measure respondents' attachment to their party, we employed the aforementioned question that is used widely within the literature and prominent studies of the political attitudes of the American public. To measure overall level of news interest, respondents were asked the following question: "Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs ..." (Most of the time, Some of the time, Only Now and Then, Hardly at all, Don't Know). The wording of this question was adopted verbatim from the American National Election Study.

To measure frequency of Twitter use, we asked respondents the following question: "How often do you visit

Twitter to read posts from other accounts?" We then created a binary variable that describes whether people visited Twitter multiple times every day, or less than once a day. We also considered using the number of accounts respondents followed on Twitter as a criterion for overall frequency of Twitter activity, but this measure was highly correlated with the self-reported measure in our survey just described and does not necessarily reflect the regularity with which a user visits Twitter.

2.7 Covariate Balance Check

The table below reports the results of t tests that show no significant difference in covariates between respondents who were assigned to the treatment condition and those who were not.

Table 2: Covariate Balance

Variable	Assigned to Treatment	Assigned to Control	p-value
Geography (Northeast)	855	636	0.14
Geography (South)	933	796	0.42
Geography (North Central)	845	642	0.78
Geography (West)	845	660	0.45
Caucasian	1158	1126	0.46
Male	995	840	0.14
Over 35	848	627	0.16

Because we ran separate experiments for Democratic and Republican respondents, we also ran separate covariate balance checks by party identification. These tests revealed balance on almost all of our control variables, apart from the dummy variable describing respondent location in the Northeastern United States (for Republicans), and gender (for Democrats). Although neither of these variables have been shown to have a particularly strong relationship with *change* in partisan identification by previous studies, we control for them in the models described below, and ran separate models that exclude these variables, which produced results that are nearly identical to those we report in the main text of our article.

Table 3: Covariate Balance Republican Experiment

Variable	Assigned to Treatment	Assigned to Control	p-value
Geography (Northeast)	373	263	0.02
Geography (South)	420	363	0.79
Geography (North Central)	363	281	0.52
Geography (West)	363	288	0.21
Caucasian	523	518	0.98
Male	438	386	0.99
Over 35	371	264	0.05

Table 4: Covariate Balance Democratic Experiment

Variable	Assigned to Treatment	Assigned to Control	p-value
Geography (Northeast)	482	373	1.00
Geography (South)	513	433	0.39
Geography (North Central)	482	361	0.37
Geography (West)	482	372	0.94
Caucasian	635	608	0.38
Male	557	454	0.05

Variable	Assigned to Treatment	Assigned to Control	p-value
Over 35	477	363	0.85

3 Treatment Delivery and Compliance

Respondents assigned to the treatment condition were invited to follow the bots on October 21, 2017, roughly one week after they were recruited to complete the pre-treatment survey. The one week buffer between the initial survey and treatment was intended to decrease the likelihood that respondents would become aware of the purpose of the experiment (see additional discussion of experiment effects below). The invitation to those in the treatment condition was as follows: "Recently you completed a survey for YouGov about how often you use Twitter. You have been randomly selected for an opportunity to receive up to 10,000 points for completing an additional task related to that survey. Participation in this portion of the study will involve following a Twitter account created as a part of the study that will tweet 24 messages per day for one month." The invitation then stated that respondents must follow the bot for an entire month in order to receive compensation and provided details about how respondents could receive additional incentives for correctly answering questions about the bot's tweets (as we describe in further detail below). Finally, the invitation clarified that participation in this portion of the study was entirely voluntary, and told respondents that they had the right to decline to participate or stop participating at any point. The final paragraph of the invitation invited them to contact the Human Subjects Committee at the first author's institution if they have any questions regarding their rights as a research participant.

If respondents accepted the invitation, they were redirected to another web page that included a link to follow either the study's liberal or conservative bot, depending upon their self-reported party identification. This page informed them they would earn the equivalent of $11 for following the bot and up to an additional $18 for successfully answering questions about the content of the messages retweeted by the bot during surveys that would follow each week. Each bot was given a non-descript name that did not prime the political ideology of opinion leaders that it retweeted. We are unable to report the names here because data collection continues for follow-up research and disclosure of the Twitter handles could be used to identify respondents in the study who engaged with the bot by commenting on, liking, or retweeting its retweets. For the first few days in which respondents began following the bots, only pictures of nature landscapes were retweeted in order to further mask the purpose of the study.

3.1 Ethics and Protection of Human Subjects

Our research was approved by the Institutional Review Boards at Duke University and New York University. All respondents digitally signed an informed consent dialogue before they participated in our research. Because open-ended questions in our pilot study indicated Republican respondents might have anti-intellectual sentiment that could create measurement error, the informed consent dialogue did not state that the research was being conducted by academic researchers, though the name of the first author's university was listed in the penultimate paragraph alongside instructions about how to contact the Institutional Review Board with complaints about the research (see informed consent screenshot above for exact language used). No such complaints were received.

Though most Twitter data is publicly available for academic research, our study links such data to confidential survey data. Because such data are highly sensitive, we do not publicly release the names, Twitter handles, or numeric ids of any respondents in our study. Nor do we make the content of their tweets, the names of the people they follow, or the names of the people who follow them publicly available. Instead, the public release of our data includes a variable that describes each respondent's a) number of Twitter followers, b) the number of people they follow, and c) a measure of the ideological heterogeneity of their Twitter network which is described in additional detail below. We coarsen these variables into increments of 50 within the public release data in order to prevent them being used to identify respondents in our study. Each member of

the research team who analyzed the non-public Twitter data signed a confidentiality pledge in accordance with evolving standards in the field of social media research.

3.2 Algorithmic Confounding

A unique property of social media research, as discussed by Lazer et al (2014), is algorithmic confounding—or the possibility that software that governs users' experience on social media websites can shape research findings. In our case, we were concerned that Twitter's "timeline" algorithm would create inconsistencies in treatment across subjects. This algorithm sorts the order in which messages appear in a user's Twitter feed, which we feared could shape exposure to our bots, since Twitter's algorithm prioritizes accounts with which a user regularly engages (either by retweeting messages from such accounts, commenting on messages produced by such accounts, or "liking" messages produced by such accounts).

To mitigate algorithmic confounding, we asked all respondents to disable Twitter's timeline algorithm in order to ensure that they viewed tweets from their bot, and were thus able to answer questions about the content of its tweets. We provided step-by-step instructions about how to disable the algorithm in our recruitment dialogue.

We also took steps to mitigate bias from Twitter's "recommender" algorithm. Our concern was that people who follow one of our study's bots may subequently receive recommendations to follow similar types of people (i.e. out-partisans), which could effectively strengthen our treatment (if the user acts upon such recommendations). Though Twitter does not make details about its recommender algorithm publicly available, Gupta et al (2013), who were involved in the design of the algorithm, indicate it makes recommendations using second-order network relationships—or who the person followed follows. Because our bots did not follow any other Twitter accounts, we believe the likelihood of bias created by the recommender algorithm was very low.

3.3 Measuring Compliance

We measured compliance with treatment assignment in several ways described in our pre-registration statement. First, we wrote code to monitor whether respondents followed one of our bots for the entire study period, and also to collect supplemental social network data that we will analyze in a subsequent study. Because it is possible that some respondents followed one of our bots but subsequently "muted" it—or changed the settings for their Twitter account so that they continue to follow one of our bots but do not receive its messages within their timeline—we employed deception. In our recruitment dialogue, we informed respondents that the research team would monitor whether they muted the account using a computer program, even though it is not possible to do so. This deception was approved by the Institutional Review Board that monitored our research and subjects were debriefed about this deception following the conclusion of our study.

Because we could not be certain that respondents who followed one of our study's bots were consistently exposed to the messages it retweeted, we took two additional steps. First, we conducted weekly surveys of respondents which asked them to answer questions about the content of our Twitter bots' messages during the previous week. In the informed consent dialogue of the pre-treatment survey, partipants were informed that participation in these weekly surveys was optional, and that participation in each survey would be compensated with up to 4,000 points, in addition to compensation for following the Twitter account for the study period.

Respondents in the treatment condition were recruited to complete weekly surveys as follows. First, they were directed to a web page within the YouGov system with the following invitation at the top of the page: "Recently we asked you to follow the [@name_of_bot_removed] Twitter account. We'd like to ask you a few questions about what you've seen posted on this account. If you answer the question correctly, YouGov will award you up to 4,000 points!" The same web page featured eight animal pictures each week at the bottom of the screen, and asked respondents, "First, which one of these pictures did [@name_of_bot_removed] retweet each day over the past week." Though our bots retweeted a picture of a cute animal twice each day, these

pictures were deleted from the bot accounts immediately prior to the distribution of each weekly survey. We developed this measure in order to determine whether respondents eyeballs passed across messages produced by our bots.

After being asked to identify the cute animal picture, respondents were redirected to a second page in which they were asked a question about the substantive content of one of the messages retweeted by our bots in the previous 72 hours. This measure was designed to gauge not only whether respondents had viewed the bots tweets, but whether they had read them as well. We designed these questions so as not to favor those with higher political knowledge or prime partisan sentiment—and also so that they could not be easily answered using a browser's "search" function. At the same time—unlike the identification of the animal pictures just described— respondents were able to revisit the bot's Twitter feed in order to find the answer to the question. We cannot list the exact wording of these questions because such information could be used to identify the name of our Twitter bots, and hence, the names of those in the study who interacted with its messages throughout the study period. Instead, we provide an example of the type of question we asked as follows: "Over the past three days, the [name of study's Twitter account here] retweeted a message about a philanthropist who gave a large amount of money to help people recover from a major disaster. How much money did this person donate?"

Figure 6 provides an overview of the number of respondents who were able to answer our questions about the bot during each of the three weekly surveys during the study period. The three weekly surveys were conducted between October 27-29th, November 3-5, and November 10-13th. The final post-treatment survey, which contained the exact same questions as the pre-treatment survey, was administered from November 23rd to November 27th.

The code below calculates the compliance rate by party in three ways. First, we calculate the percentage of respondents assigned to the treatment condition within each party who accepted our invitation to follow one of the study's bots and did so for the entire study period. Second we create a six-point compliance scale that describes the number of questions that respondents were able to answer correctly during the three "compliance check" surveys administered each week during the study period.

In the models below, we employ three dichotomous variables to measure compliance. The first describes individuals in the treatment condition who followed the bot for the entire study period (what we call "minimal compliance" in the main text of our article). The second describes respondents who answered more than one question correctly but not all questions correctly (what we call "partial compliance" in the main text of our article). The third describes respondents who answered all questions asked in the weekly surveys correctly (what we call "full compliance" in the main text of our article).

```
#calculate compliance rate by party identification
democrats<-twitter_data[twitter_data$party_id_wave_1==1,]
nrow(democrats[democrats$bot_followers==1,]) / nrow(democrats[democrats$treat==1,])
republicans<-twitter_data[twitter_data$party_id_wave_1==2,]
nrow(republicans[republicans$bot_followers==1,]) / nrow(republicans[republicans$treat==1,])

#construct continuous compliance measure
twitter_data$complier_scale<-0
twitter_data$complier_scale <-
  rowSums(twitter_data[,c("substantive_question_correct_wave_2",
  "substantive_question_correct_wave_3",
  "substantive_question_correct_wave_4",
  "animal_correct_wave_2",
  "animal_correct_wave_3",
  "animal_correct_wave_4")], na.rm=TRUE)


#construct partial complier dummy
```

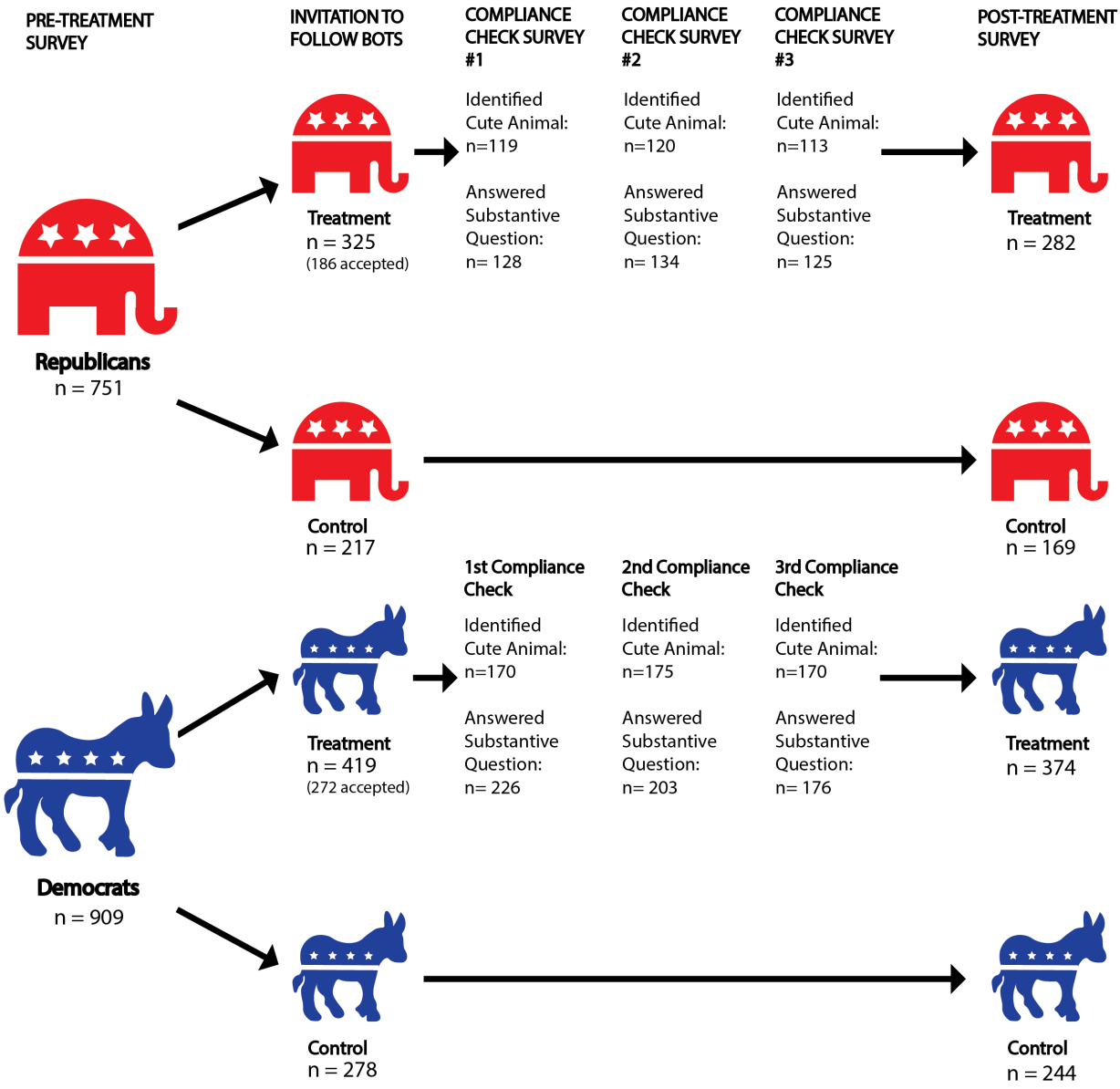


Figure 6: Response to Five Surveys Conducted During Survey Period

```
twitter_data$half_complier<-0
twitter_data$half_complier[twitter_data$complier_scale>0&
                           twitter_data$complier_scale<6]<-1

#construct full complier dummy
twitter_data$perfect_complier<-0
twitter_data$perfect_complier[twitter_data$complier_scale==6]<-1


#calculate percentage of respondents who answered all substantive questions correctly
nrow(twitter_data[twitter_data$substantive_question_correct_wave_2==1&
                  twitter_data$substantive_question_correct_wave_3==1&
                  twitter_data$substantive_question_correct_wave_4==1,])/
       nrow(twitter_data[twitter_data$bot_followers==1,])

#calculate percentage of respondents who identified all animal pictures correctly
nrow(twitter_data[twitter_data$animal_correct_wave_2==1&
                  twitter_data$animal_correct_wave_3==1&
                  twitter_data$animal_correct_wave_4==1,])/
       nrow(twitter_data[twitter_data$bot_followers==1,])
```

3.4 Design of Twitter Bots

The two Twitter bots we created for this study were designed as follows. First, we built upon Barbera et al.'s (2015) ideological scoring method for Twitter users. We began by collecting the Twitter IDs, or "handles," for all presidential candidates and members of the House and Senate as of August 5, 2017. We then scraped the names of all people who these elected officials follow from Twitter's Application Programming Interface, which yielded a total sample of 636,737 Twitter accounts. Next, we eliminated all those who were not followed by at least 15 of the aforementioned elected officials. We then conducted a correspondence analysis on the resultant adjacency matrix, and used the first principal component to create a liberal/conservative score for all of those in this "opinion leader" network. We binned this scale into seven quantiles, and dropped those in the fourth (centrist) quantile. The liberal bot randomly retweeted messages from opinion leaders in the first, second, and third quantile produced during the preceeding 24 hours, and the conservative bot randomly retweeted messages from opinion leaders in the fifth, six, and seventh quantiles during the preceeding 24 hours.

We took several additional steps to improve the ideological scores we used to create our bots. First, we eliminated all U.S. government agencies, since most of these retweeted non-partisan messages that would dilute our treatment. Second, we eliminated all accounts that were administered by for-profit U.S. corporations, though we did not eliminate non-profit organizations, think tanks, or other nonprofit groups. Third, we eliminated a small number of accounts that were controlled by elected officials outside the United States.

Despite these steps, pilot analyses of the ideological continuum consistently identified a small number of elected officials who were misclassified according to our measure. Each of these individuals were very high profile opinion leaders such as Mitch McConnell and John McCain, who have very large followings that include a large number of non-Republicans, which made them centrists instead of conservatives in our original analysis. We thus reclassified the small number of elected officials who were mistakenly identified by assigning them a random ideological score between the first and second quantile of opinion leaders that defined their party using the first principal component measure described above.

The liberal and conservative Twitter bots created by our research team were both hosted on an Amazon EC2 server. Every hour, our program randomly drew a message produced by an elected official or opinion leader from the previous 24 hours from one of the two samples. During three of the four weeks during the study period, the bot retweeted a different animal picture at two random times each day, as we described above.

To further illustrate the types of Twitter accounts retweeted by the study's bots, Table 5 lists the words that appear most frequently in the "about me," or biographical section, of the Twitter accounts retweeted by both the liberal and conservative bots. Similarly, Table 6 lists the top words that appeared in messages retweeted by the two bots during the study period.

Table 5: Top Words from Biographies of Twitter Accounts Retweeted by Bots

Liberal Accounts	Freq.	Conservative Accounts	Freq.
news	238	district	167
reporter	184	news	155
political	124	u.s	119
politics	121	official	103
editor	120	congressional	97
national	120	proudly	79
house	105	representing	79
correspondent	103	house	78
washington	103	twitter	78
politico	91	account	74
author	86	conservative	74
cnn	84	husband	69
official	83	author	58
u.s	77	chairman	58
covering	75	political	57
twitter	73	serving	57
senior	72	host	56
district	69	father	55
policy	68	congressman	53
host	66	congress	50
white	62	national	49
congress	61	senator	49
congressional	58	fox	47
people	57	people	46
writer	57	politics	46
account	56	proud	45
fan	55	american	44
chief	54	http	44
endorsements	54	represent	44
tweets	53	policy	43
alum	52	tweets	43
health	52	editor	40
email	51	follow	40
american	49	republican	39
world	48	committee	38
msnbc	45	texas	38
post	45	media	37
representing	45	contributor	36
times	44	endorsement	34
endorsement	41	representatives	34
follow	41	business	33
columnist	40	united	33
tips	40	founder	32
public	38	freedom	32
husband	37	writer	32

Table 6: Top Words from Messages Retweeted by Bots

Liberal Bot's Tweets	Freq.	Conservative Bot's Tweets	Freq.
trump	256	tax	125
tax	93	trump	123
people	85	day	86
time	74	house	84
house	64	time	78
president	60	support	65
day	56	people	64
congress	49	read	62
bill	48	news	61
daca	48	reform	56
gop	48	national	53
health	48	watch	52
white	48	join	50
trump's	43	congress	49
week	43	harvey	48
hurricane	41	bill	47
fight	40	http	47
news	40	taxreform	47
read	40	american	46
story	40	act	45
it's	38	senate	45
talk	38	texas	44
harvey	37	jobs	42
americans	36	america	41
senate	36	happy	41
join	34	week	41
t.c	34	proud	40
u.s	34	live	39
world	34	president	38
change	33	learn	37
climate	31	realdonaldtrump	37
dreamers	31	potus	36
plan	31	u.s	36
vote	31	gop	35
women	31	families	34
gt	30	health	34
support	30	life	34
htt	29	hurricane	33
live	29	family	32
tomorrow	29	free	32
american	28	students	32
care	28	veterans	32
deal	28	discuss	31
love	28	foxnews	31
statement	28	americans	30

4 Outcome Measure and Controls

4.1 Creating Outcome Index

Our study examined a range of political attitudes, but in this article we focus upon shifts in what is often called "ideological polarization," or differences in attitudes about policy issues that consistently divide Democrats and Republicans such as inequality, race, and immigration. Though ideological scoring of roll-call voting has been the subject of extended analysis for some time, indices of liberalism vs. conservatism among the broader public are fewer (Bafumi and Herron 2010; Jessee 2012; Tausanovitch and Warshaw 2013; Hill and Tausanovitch 2015). We employed a variation of the "ideological consistency scale" developed by previous studies because it measures liberal vs. conservative opinions via a battery of questions in order to minimize the measurement error that might occur on a single survey item (Dimock and Carroll 2014). The scale, which asks respondents to agree or disagree with a series of twenty statements about social policies worded to favor either liberal or conservative views, was previously included in sixteen nationally representative surveys. We make two important modifications to this scale. Instead of a binary choice between liberal and conservative options for each policy statement, we use a seven-point response scale, since allowing respondents to indicate strength or extremity of opinion provides a more accurate measure of ideological polarization (Fiorina, Abrams, and Pope 2006; Hill and Tausanovitch 2015). Second, instead of asking respondents to read twenty questions, we randomly selected five liberal versions of each policy statement and five conservative versions. These modifications were important because we did not expect our intervention would completely change people's partisan positions (from Democrats to Republicans), and also because the format just described minimizes cognitive load.

Our survey asked respondents to agree or disagree with the following statements on a seven point scale from "strongly disagree" to "strongly agree."

1) "Stricter environmental laws and regulations cost too many jobs and hurt the economy."
2) "Government regulation of business is necessary to protect the public interest."
3) "Poor people today have it easy because they can get government benefits without doing anything in return."
4) "Immigrants today strengthen our country because of their hard work and talents."
5) "Government is almost always wasteful and inefficient."
6) "The best way to ensure peace is through military strength."
7) "Racial discrimination is the main reason why many black people can't get ahead these days."
8) "The government today can't afford to do much more to help the needy."
9) "Business corporations make too much profit."
10) "Homosexuality should be accepted by society."

As we mentioned, half of these statements are worded in a manner that is designed to appeal to liberals (#2,#4,#7,#9,#10), and the other half are intended to appeal to conservatives (#1,#3,#5,#6,#8). Question order was randomized in both the pre- and post-treatment surveys.

The code below was used to create our outcome measure. Liberal questions were reverse-coded such that negative values on our outcome indicate respondents becoming more liberal and positive values indicate respondents becoming more conservative. We calculate the mean score on this ten-item index for the pre- and post-treatment survey, and our models predict the post-treatment scale score, controlling for the pre-treatment scale score.

```
#invert questions that prime liberal values
twitter_data$government_should_regulate_businesses_wave_1<-
```

```
  8-twitter_data$government_should_regulate_businesses_wave_1
twitter_data$racial_discrimination_hurts_black_people_wave_1<-
  8-twitter_data$racial_discrimination_hurts_black_people_wave_1
twitter_data$immigrants_strengthen_country_wave_1<-
  8-twitter_data$immigrants_strengthen_country_wave_1
twitter_data$corporations_make_too_much_profit_wave_1<-
  8-twitter_data$corporations_make_too_much_profit_wave_1
twitter_data$homosexuality_should_be_accepted_wave_1<-
  8-twitter_data$homosexuality_should_be_accepted_wave_1
twitter_data$government_should_regulate_businesses_wave_5<-
  8-twitter_data$government_should_regulate_businesses_wave_5
twitter_data$racial_discrimination_hurts_black_people_wave_5<-
  8-twitter_data$racial_discrimination_hurts_black_people_wave_5
twitter_data$immigrants_strengthen_country_wave_5<-
  8-twitter_data$immigrants_strengthen_country_wave_5
twitter_data$corporations_make_too_much_profit_wave_5<-
  8-twitter_data$corporations_make_too_much_profit_wave_5
twitter_data$homosexuality_should_be_accepted_wave_5<-
  8-twitter_data$homosexuality_should_be_accepted_wave_5

#calculate chronbach's alpha
alpha_calc<-twitter_data[,c(
          "government_should_regulate_businesses_wave_1",
          "racial_discrimination_hurts_black_people_wave_1",
          "immigrants_strengthen_country_wave_1",
          "corporations_make_too_much_profit_wave_1",
          "homosexuality_should_be_accepted_wave_1",
          "government_wasteful_inefficient_wave_1",
          "poor_people_have_it_easy_wave_1",
          "government_cannot_afford_to_help_needy_wave_1",
          "best_way_peace_military_strength_wave_1",
          "stricter_environmental_laws_damaging_wave_1")]
library(psych)
psych::alpha(alpha_calc)

#create average score by wave
twitter_data$substantive_ideology_scale_wave_1<-rowMeans(twitter_data[,c(
          "government_should_regulate_businesses_wave_1",
          "racial_discrimination_hurts_black_people_wave_1",
          "immigrants_strengthen_country_wave_1",
          "corporations_make_too_much_profit_wave_1",
          "homosexuality_should_be_accepted_wave_1",
          "government_wasteful_inefficient_wave_1",
          "poor_people_have_it_easy_wave_1",
          "government_cannot_afford_to_help_needy_wave_1",
          "best_way_peace_military_strength_wave_1",
          "stricter_environmental_laws_damaging_wave_1")], na.rm=TRUE)


twitter_data$substantive_ideology_scale_wave_5<-rowMeans(twitter_data[,c(
          "government_should_regulate_businesses_wave_5",
          "racial_discrimination_hurts_black_people_wave_5",
          "immigrants_strengthen_country_wave_5",
```

```
          "corporations_make_too_much_profit_wave_5",
          "homosexuality_should_be_accepted_wave_5",
          "government_wasteful_inefficient_wave_5",
          "poor_people_have_it_easy_wave_5",
          "government_cannot_afford_to_help_needy_wave_5",
          "best_way_peace_military_strength_wave_5",
          "stricter_environmental_laws_damaging_wave_5")], na.rm=TRUE)
```

4.2 Control Variables

The models described below include a variety of control variables collected from our pre-treatment survey, Twitter's Application Programming Interface, and YouGov. We also ran separate models without controls, which yielded very similar results (Republicans: t=2.74, p<.006, Democrats: t= -1.54, p<.12). We obtained standard demographic variables about all respondents (age, income, education, gender, race, and geographic region) from YouGov. Age is defined as the year in which the respondent was born. Income is coded on a sixteen point scale that ranges from "less than $10,000" to "$500,000 or more." Gender is a binary variable coded as positive if the respondent is female. The models also include binary variables that indicate whether a respondent is from one of five geographic regions defined in the American National Election Study.

We also created variables designed to measure the strength of respondents' echo chambers pre-treatment. We asked respondents a battery of questions about their media consumption practices, and requested they list the top three media sources they consume most frequently in order to determine the amount of ideological bias in their media diet pre-treatment. Unfortunately, more than 25% of respodents did not provide the names of media sources for which we could identify ideological leaning, so we were ultimately unable to include this variable in our analyses. Fortunately, we also calculated the percentage of people who the respondent follows on Twitter who share their party identification pre-treatment using the network-based ideological scoring method for Twitter users described above. This measure is highly correlated with the aforementioned media consumption measure, so we include the Twitter-based metric to measure the strength of respondents' echo chambers pre-treatment. Our pre-treatment survey also asked respondents to estimate the percentage of people in their offline networks who share their party identification in order to further capture the strength of ideological bias within their offline social networks, which have been shown to have higher ideological bias than online networks (Gentzkow and Shapiro 2011). We include this continuous measure in all of our models as well. We did not detect significant multicollinearity between these two variables. Finally, our models also include a continuous measure of the number of people the respondent follows on Twitter in order to account for the likelihood that our bot's messages may be more difficult to see for those who follow more people.

The code below recodes and subsets the control variables for analyses that we conduct below. The `bin_maker` variable describes the randomization blocks described below. This variable is defined by the `political_wave_1` variable (overall level of interest in the news), the `freq_twitter_wave_1` variable (frequency of Twitter usage), and the `strong_partisan` variable describes whether respondents described themselves as "strong" Democrats or Republicans.

```
#subset control variables and variable used for block randomization (bin_maker)
control_variables<-twitter_data[,c(
    "caseid",
    "birth_year",
    "family_income",
    "education",
    "gender",
    "ideo_homogeneity_offline",
    "northeast",
    "north_central",
    "south",
```

```
    "west",
    "percent_co_party",
    "friends_count_wave_1",
    "strong_partisan",
    "political_wave_1",
    "freq_twitter_wave_1",
    "bin_maker")]
```

4.3 Missing Data

We employed multiple imputation to address a small amount of missing data in our pre-treatment survey–particularly the variables that describe respondents' income (7% missing) as well as their estimate of the ideological composition of their offline networks (1% missing). We did not impute missing responses for the second wave outcome measure.

```
#examine missing data in first wave
library(Amelia)
missmap(control_variables)

#impute missing data from first wave
forimputation<-cbind(twitter_data$substantive_ideology_scale_wave_1, control_variables)
colnames(forimputation)[colnames(forimputation)==
              "twitter_data$substantive_ideology_scale_wave_1"]<-
                      "substantive_ideology_scale_wave_1"

#prepare variables for imputation
library(mice)
forimputation$caseid<-as.character(forimputation$caseid)
#take log of variables with heavy skew
forimputation$percent_co_party<-log(forimputation$percent_co_party+1)
forimputation$friends_count_wave_1<-log(forimputation$friends_count_wave_1+1)

#impute
imputed_data <- mice(forimputation,m=15,seed=352,
                     exclude=c("caseid","bin_maker"))
imputed_data <- complete(imputed_data,action=15)

#reassemble dataaset with additional variables we need for subsequent analysis
to_bind<-twitter_data[,c("treat",
                  "perfect_complier",
                  "half_complier",
                  "bot_followers",
                  "party_id_wave_1",
                  "party_strength_wave_1",
                  "substantive_ideology_scale_wave_5",
                  "endtime_wave_5")]

final_data<- cbind(to_bind, imputed_data)


save(final_data, file="Final Data for Models.Rdata")
```

5 Calculating Treatment Effects

To evaluate the effect of our intervention, we calculated both Intent-to-Treat (ITT) effects as well as Complier Average Causal Effects (CACE) that account for level of treatment compliance for experiments with both Democrats and Republicans. These models predict the post-treatment score on our ten-item liberal/conservative scale controlling for respondents' pre-treatment score on this scale, twelve additional covariates (described above), as well as a factor variable (`bin_maker`) that describes our treatment blocks. The models for Republicans and Democrats are reported in Tables 6 and 7 below. At the suggestion of an anonymous reviewer, we also ran an additional pooled model that combined both Republicans and Democrats into a single model where the outcome was absolute change in attitudes and the treatment indicator was interacted with a binary variable that described whether the respondent was a Republican. This model produced very similar results, but we do not report them here because of concerns about scaling issues and because the two experiments were conducted in isolation from each other.

5.1 Intent-to-Treat Effects

```
#subset Republicans and Democrats and drop missing data from post-treatment survey
republicans <- final_data[final_data$party_id_wave_1==2,]
democrats <- final_data[final_data$party_id_wave_1==1,]

#Republicans
republican_ITT_model<-lm(substantive_ideology_scale_wave_5~
                      #treatment assignment variable
                      treat+
                      #pre-treatment ideology score
                      substantive_ideology_scale_wave_1+
                      #% of people followed on Twitter from same party
                      percent_co_party+
                      #% of people in offline networks from same party
                      ideo_homogeneity_offline+
                      #total number of people followed pre-treatment
                      friends_count_wave_1+
                      #demographics
                      birth_year +
                      family_income+
                      education+
                      gender+
                      northeast+
                      north_central+
                      south+
                      #factor variable used to create treatment blocks
                      as.factor(bin_maker),
                      data=republicans)

#calculate robust standard errors
library(lmtest)
library(sandwich)

coefficients<-as.data.frame(coeftest(republican_ITT_model,
                                     vcov = vcovHC(republican_ITT_model, type="HC1"))[2:13,1:4])

library(pander)
```

```
panderOptions('digits',3)
panderOptions('table.split.table', 300)
set.caption("Intent-to-Treat Model (Republicans)")
pander(coefficients)
```

Table 7: Intent-to-Treat Model (Republicans)

	Estimate	Std. Error	t value	Pr(>|t|)
treat	0.12	0.0442	2.72	0.00681
substantive_ideology_scale_wave_1	0.819	0.0279	29.3	2.27e-103
percent_co_party	0.227	0.109	2.08	0.0385
ideo_homogeneity_offline	0.00173	0.00109	1.59	0.112
friends_count_wave_1	0.042	0.0146	2.86	0.00438
birth_year	-5.11e-05	0.00161	-0.0316	0.975
family_income	0.00626	0.00738	0.849	0.396
education	-0.00593	0.0164	-0.361	0.718
gender	-0.0277	0.0454	-0.611	0.542
northeast	0.0211	0.0794	0.266	0.791
north_central	-0.0873	0.066	-1.32	0.186
south	-0.0582	0.0551	-1.06	0.291

```
#Democrats
democrat_ITT_model<-lm(substantive_ideology_scale_wave_5~
                       #treatment assignment variable
                       treat+
                       #pre-treatment ideology score
                       substantive_ideology_scale_wave_1+
                       #% of people followed on Twitter from same party
                       percent_co_party+
                       #% of people in offline networks from same party
                       ideo_homogeneity_offline+
                       #total number of people followed pre-treatment
                       friends_count_wave_1+
                       #demographics
                       birth_year +
                       family_income+
                       education+
                       gender+
                       northeast+
                       north_central+
                       south+
                       #factor variable used to create treatment blocks
                       as.factor(bin_maker),
                       data=democrats)

#calculate robust standard errors
library(lmtest)
library(sandwich)

coefficients<-as.data.frame(coeftest(democrat_ITT_model,
                                     vcov = vcovHC(democrat_ITT_model, type="HC1"))[2:13,1:4])
```

```
library(pander)
panderOptions('digits',3)
panderOptions('table.split.table', 300)
set.caption("Intent-to-Treat Model (Democrats)")
pander(coefficients)
```

Table 8: Intent-to-Treat Model (Democrats)

	Estimate	Std. Error	t value	Pr(>|t|)
treat	-0.0248	0.0335	-0.739	0.46
substantive_ideology_scale_wave_1	0.834	0.0229	36.5	1.14e-153
percent_co_party	-0.109	0.101	-1.08	0.28
ideo_homogeneity_offline	0.000554	0.000718	0.771	0.441
friends_count_wave_1	-0.000722	0.0122	-0.0591	0.953
birth_year	-0.00107	0.00115	-0.934	0.351
family_income	0.00158	0.00492	0.322	0.747
education	0.00212	0.0145	0.146	0.884
gender	0.031	0.0346	0.893	0.372
northeast	-0.0497	0.0486	-1.02	0.307
north_central	0.0194	0.0447	0.435	0.664
south	-0.0152	0.0425	-0.358	0.72

5.2 Complier Average Causal Effects

We calculated the Complier Average Causal Effect (CACE) using the two-stage least squares approach developed by Imbens and Rubin (2015). In the models reported in the main text of our manuscript, we report results for respondents who were fully compliant (indicated by answering all weekly compliance checks correctly), partially compliant (those who answered more than one of the weekly compliance check questions correctly but less than six), and minimally compliant (those who followed one of our study's bots for the entire study period).

The following assumptions are required to estimate CACE: 1) Ignorability, 2) Monotonicity, 3) Stable Unit Treatment Value, 4) Non-Interference, and, 5) Excludability. The first and second of these assumptions are supported by our research design, and we believe the third and fourth assumptions are warranted because of the extensive steps we took to eliminate respondents from the initial pre-treatment survey who followed each other on Twitter. The excludability assumption—or the assumption that those who did not comply with treatment have the same potential outcomes as those in control—is more problematic. We believe this assumption is warranted for our most basic definition of compliance: whether or not respondents assigned to treatment accepted our invitation to follow the bot. Because we were able to monitor who was following the bot at all times, it is unlikely that anyone who was invited to follow the bot but did not do so was ultimately exposed to its messages—particularly in light of our aforementioned attempts to mitigate causal interference. On the other hand, the excludability assumption is arguably less reasonable for our two other compliance measures, which describe whether respondents who followed the bot were able to answer some or all of the questions our surveys asked them about the bot's tweets. This is because it is likely that some of the respondents who accepted our invitation to follow the bot but did not answer any questions correctly about the content of its tweets were nevertheless exposed to some of its messages. Because this assumption is rather strong, we provide multiple estimates of our treatment effects (see above), and encourage readers to focus on our most basic "minimal" compliance measure (whether or not respondents' accepted our invitation to follow one of the study's two bots)

The code below was used to calculate CACE for the three different levels of compliance described above:

```
#create list of datasets
datasets<-list(democrats, republicans)

#create function to calculate Complier Average Causal effect for
#fully compliant respondents
library(ivpack)
CACE_fc<-function(data){
  #drop cases without outcome response for final survey
  data<-data[complete.cases(data),]
  results<-ivreg(substantive_ideology_scale_wave_5 ~
                       perfect_complier+
                       substantive_ideology_scale_wave_1+
                       percent_co_party+
                       friends_count_wave_1+
                       birth_year +
                       family_income+
                       education+
                       gender+
                       ideo_homogeneity_offline+
                       northeast+
                       north_central+
                       south+
                       as.factor(bin_maker)
                       |
                       treat+
                       substantive_ideology_scale_wave_1+
                       percent_co_party+
                       friends_count_wave_1+
                       birth_year+
                       family_income+
                       education+
                       gender+
                       ideo_homogeneity_offline+
                       northeast+
                       north_central+
                       south+
                       as.factor(bin_maker),
                       data = data)
      #calculate robust standard errors
      output<-robust.se(results)[2,]
      return(output)}


#create function to calculate Complier Average Causal Effect for
#partially compliant respondents
CACE_hc<-function(data){
  #drop cases without outcome response for final survey
  data<-data[complete.cases(data),]
  results<-ivreg(substantive_ideology_scale_wave_5 ~
                       half_complier+
                       substantive_ideology_scale_wave_1+
                       percent_co_party+
                       friends_count_wave_1+
```

```
                      birth_year +
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker)
                      |
                      treat+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year+
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker),
                      data = data)
      #calculate robust standard errors
      output<-robust.se(results)[2,]
      return(output)}


#create function to calculate Complier Average Causal Effect for
#respondents who followed bot
CACE_bf<-function(data){
  #drop cases without outcome response for final survey
  data<-data[complete.cases(data),]
  results<-ivreg(substantive_ideology_scale_wave_5 ~
                      bot_followers+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year +
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker)
                      |
                      treat+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
```

```
                    birth_year+
                    family_income+
                    education+
                    gender+
                    ideo_homogeneity_offline+
                    northeast+
                    north_central+
                    south+
                    as.factor(bin_maker),
                    data = data)
      #calculate robust standard errors
      output<-robust.se(results)[2,]
      return(output)}



#run models
full_compliance_models <- lapply(datasets, function(x) CACE_fc(x))
half_compliance_models <- lapply(datasets, function(x) CACE_hc(x))
bot_follower_models <- lapply(datasets, function(x) CACE_bf(x))

#extract results for republicans
republican_full_compliance_cace<-as.data.frame(t(full_compliance_models[[2]]))
republican_full_compliance_cace$sample<-"republicans_full_compliance"
republican_full_compliance_cace$party<-"republicans"
names(republican_full_compliance_cace)<-c("estimate","se","t","p","sample","party")
republican_half_compliance_cace<-as.data.frame(t(half_compliance_models[[2]]))
republican_half_compliance_cace$sample<-"republicans_half_compliance"
republican_half_compliance_cace$party<-"republicans"
names(republican_half_compliance_cace)<-c("estimate","se","t","p","sample","party")
republican_bot_follower_cace<-as.data.frame(t(bot_follower_models[[2]]))
republican_bot_follower_cace$sample<-"republicans_bot_follower"
republican_bot_follower_cace$party<-"republicans"
names(republican_bot_follower_cace)<-c("estimate","se","t","p","sample","party")

#extract results for democrats
democrat_full_compliance_cace<-data.frame(t(full_compliance_models[[1]]))
democrat_full_compliance_cace$sample<-"democrats_full_compliance"
democrat_full_compliance_cace$party<-"democrats"
names(democrat_full_compliance_cace)<-c("estimate","se","t","p","sample","party")
democrat_half_compliance_cace<-as.data.frame(t(half_compliance_models[[1]]))
democrat_half_compliance_cace$sample<-"democrats_half_compliance"
democrat_half_compliance_cace$party<-"democrats"
names(democrat_half_compliance_cace)<-c("estimate","se","t","p","sample","party")
democrat_bot_follower_cace<-as.data.frame(t(bot_follower_models[[1]]))
democrat_bot_follower_cace$sample<-"democrats_bot_follower"
democrat_bot_follower_cace$party<-"democrats"
names(democrat_bot_follower_cace)<-c("estimate","se","t","p","sample","party")
```

The following code was used to produce Figure 3 in the main text of our article.

```
#create another dataset that combines ITT and CACE results for plotting
republican_itt<-
```

```
  data.frame(t(summary(republican_ITT_model, cluster="bin_maker")$coefficients[2:2,]))
names(republican_itt)<-c("estimate","se","t","p")
republican_itt$sample<-"republicans_itt"
republican_itt$party<-"republicans"

democrat_itt<-
  data.frame(t(summary(democrat_ITT_model, cluster="bin_maker")$coefficients[2:2,]))
names(democrat_itt)<-c("estimate","se","t","p")
democrat_itt$sample<-"democrats_itt"
democrat_itt$party<-"democrats"


republican_plot<-rbind(republican_full_compliance_cace,
                       republican_half_compliance_cace,
                       republican_bot_follower_cace,
                       republican_itt)

democrat_plot<-rbind(democrat_full_compliance_cace,
                     democrat_half_compliance_cace,
                     democrat_bot_follower_cace,
                     democrat_itt)


#calculate N of compliance to add to plot

republicans<-final_data[final_data$party_id_wave_1==2,]
democrats<-final_data[final_data$party_id_wave_1==1,]
nrow(republicans[republicans$perfect_complier==1,])
nrow(republicans[republicans$half_complier==1,])
nrow(republicans[republicans$bot_followers==1,])
nrow(republicans[republicans$treat==1,])
nrow(democrats[democrats$perfect_complier==1,])
nrow(democrats[democrats$half_complier==1,])
nrow(democrats[democrats$bot_followers==1,])
nrow(democrats[democrats$treat==1,])



republican_plot$sample<-factor(republican_plot$sample,
                               levels=c("republicans_full_compliance",
                                 "republicans_half_compliance",
                                 "republicans_bot_follower",
                                 "republicans_itt"),
                               labels=c("Fully Compliant Respondents (n=53)",
                                  "Partially Compliant Respondents (n=121)",
                                  "Minimally Compliant Respondents (n=181)",
                                  "Respondents Assigned to Treatment (n=316)"))

democrat_plot$sample<-factor(democrat_plot$sample,
                               levels=c("democrats_full_compliance",
                                 "democrats_half_compliance",
                                 "democrats_bot_follower",
                                 "democrats_itt"),
```

```
                                labels=c("Fully Compliant Respondents (n=66)",
                                   "Partially Compliant Respondents (n=211)",
                                   "Minimally Compliant Respondents (n=271)",
                                   "Respondents Assigned to Treatment (n=416)"))


#create standard error bars
interval1 <- -qnorm((1-0.9)/2)  # 90% multiplier
interval2 <- -qnorm((1-0.95)/2)  # 95% multiplier

#create plot
library(ggplot2)
figure_3_dems<-ggplot(democrat_plot)+
  geom_hline(yintercept = 0, colour = gray(1/2), lty = 2)+
  geom_point(aes(x=sample, y=estimate),
             position = position_dodge(width = 1/2),
             size=2, colour="blue")+
  geom_linerange(aes(x = sample, ymin = estimate - se*interval1,
                     ymax = estimate + se*interval1),
                 lwd = 1, position = position_dodge(width = 1/2),
                 colour="blue")+
  geom_linerange(aes(x = sample, y = estimate, ymin = estimate - se*interval2,
                     ymax = estimate + se*interval2),
                 lwd = .5, position = position_dodge(width = 1/2),
                 colour="blue")+
  theme(axis.text=element_text(size=12, face="bold",colour="black"),
        plot.title = element_text(face="bold", size=16, hjust = 0.5,vjust=3),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background=element_blank(),
        axis.title=element_text(size=12, colour="black"),
        legend.position="none",
        legend.key = element_blank(),
        legend.title=element_blank())+
  ylim(c(-1,1))+
  labs(x="",y="")+
  coord_flip()+
  ggtitle("Democrats")


figure_3_reps<-ggplot(republican_plot)+
  geom_hline(yintercept = 0, colour = gray(1/2), lty = 2)+
  geom_point(aes(x=sample, y=estimate),
             position = position_dodge(width = 1/2),
             size=2, colour="red")+
  geom_linerange(aes(x = sample, ymin = estimate - se*interval1,
                     ymax = estimate + se*interval1),
                 lwd = 1, position = position_dodge(width = 1/2),
                 colour="red")+
  geom_linerange(aes(x = sample, y = estimate, ymin = estimate - se*interval2,
                     ymax = estimate + se*interval2),
                 lwd = .5, position = position_dodge(width = 1/2),
                 colour="red")+
```

```
  theme(axis.text=element_text(size=12, face="bold",colour="black"),
        plot.title = element_text(face="bold", size=16, hjust = 0.5, vjust=3),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.background=element_blank(),
        axis.title=element_text(size=12, colour="black"),
        legend.position="none",
        legend.key = element_blank(),
        legend.title=element_blank())+
  labs(x="",y="")+
  ylim(c(-1,1))+
  coord_flip()+
  ggtitle("Republicans")

ggsave(figure_3_dems, file="Figure 3 dems.png", width=7, height=4, dpi=1000)
ggsave(figure_3_reps, file="Figure 3 reps.png", width=7, height=4, dpi=1000)
```

5.3 Intent-to-Treat Effects without Covariates

In order to calculate the most basic, or "pure," estimate of our causal effect, we also calculated Intent-to-Treat effects without covariates using Fisher's exact score to account for the variables we used to define treatment blocks. According to this analysis, Republicans assigned to treatment increased .10 points on our liberal-conservative scale (p<.01), where higher values indicate increased conservatism. Democrats, for their part, decreased .03 points (p<.84)

```
#create function to calculate fisher's exact score
#with grouping variable to account for blocking
fisher.exact <- function(Yobs,Wobs,Gobs){
  #Yobs is the observed outcome
  #Wobs is the observed randomziation (T/C)
  #Gobs is the group assignment
  tmp <- which(is.na(Yobs) | is.na(Wobs) | is.na(Gobs))
  Yobs <- Yobs[-tmp]
  Wobs <- Wobs[-tmp]
  Gobs <- Gobs[-tmp]
  J <- 1000
  Yfull <- matrix(c(Yobs,Yobs),nrow=length(Yobs),ncol=2, byrow=FALSE)
  tmp <- NULL
  for(i in 1:J){
    Wtmp <- rep(0,length(Yobs))
    track_means_treat <- track_means_control <- bin_size <- NULL
    for(g in unique(Gobs)){
      Wtmp[Gobs==g] <- c(rep(0,sum(1-Wobs[Gobs==g])),
                         rep(1,sum(Wobs[Gobs==g])))[sample(1:sum(Gobs==g))]
      track_means_treat <- c(track_means_treat,mean(Yfull[Wtmp==1 | Gobs==g,2]))
      track_means_control <- c(track_means_control,mean(Yfull[Wtmp==0 | Gobs==g,1]))
      bin_size <- c(bin_size,sum(Gobs==g))
    }
    tmp <- c(tmp,weighted.mean(track_means_treat,bin_size)-
               weighted.mean(track_means_control,bin_size))
  }
  track_means_treat <- track_means_control <- bin_size <- NULL
```

```
  for(g in unique(Gobs)){
    track_means_treat <- c(track_means_treat,mean(Yfull[Wobs==1 | Gobs==g,2]))
    track_means_control <- c(track_means_control,mean(Yfull[Wobs==0 | Gobs==g,1]))
    bin_size <- c(bin_size,sum(Gobs==g))
  }
  tobs = weighted.mean(track_means_treat,bin_size)-
    weighted.mean(track_means_control,bin_size)
  list(null = tmp,tobs = tobs)
}

#subset Republicans and Democrats
republicans <- final_data[final_data$party_id_wave_1==2,]
democrats <- final_data[final_data$party_id_wave_1==1,]

#Republicans
fisher_exact<-fisher.exact(republicans$substantive_ideology_change,
                           republicans$treat,
                           republicans$bin_maker)

#ATE
republican_itt<-fisher_exact$tobs

#p-Value
republican_itt_p_value<-mean(fisher_exact$null>fisher_exact$tobs)


#Democrats
fisher_exact<-fisher.exact(democrats$substantive_ideology_change,
                           democrats$treat,
                           democrats$bin_maker)

#ATE
democrat_itt<-fisher_exact$tobs

#p-Value
democrat_itt_p_value<-mean(fisher_exact$null>fisher_exact$tobs)
```

5.4 Interpretation of Effect Size

To further evaluate the effect sizes reported in the main text of our article, we offer an approximate comparison of our findings to historical shifts in the liberal/conservative scale described above, which has been administered sixteen times to a representative sample of American adults between 1994 and 2014 (Dimock and Carroll 2014). Our results are not directly comparable to these previous surveys for several reasons. Previous studies asked respondents ten questions about a social policy issue, each of which had a liberal and conservative-leaning statements, and the respondent was asked to place themselves on a continuum between the two. In contrast, we randomly selected five of these liberal statements and five conservative statements and then asked respondents to agree or disagree with them on a seven-point scale. These divergent scales result in different variance structures which makes a direct comparison of the effects impossible.

Table 9 presents data from Dimock and Carroll (2014) that aggregates the ideological consistency score into six categories that describe the percentage of the American population that is liberal or conservative. These results were created by combining binary responses into a ten-point liberal (-10) to conservative (+10) scale and then examining the distribution of responses within six quantiles.

Table 9: Distribution of Liberal/Conservative Index Over Time (Dimock and Carroll 2014)

% Who are	1994	1999	2004	2011	2014
Consistently Conservative	7	4	3	7	9
Mostly Conservative	23	16	15	19	18
Mixed	49	49	49	49	42
Mostly Liberal	18	25	25	23	22
Consistently Liberal	3	6	8	8	12
Mean Score (-10 to 10)	.6	-.6	-.9	-.3	-.6

To create an approximate comparison of the size of the conservative backfire effect for fully compliant respondents that we report in our main paper (unstandardized beta=.60) to these data, one can convert the former into a twenty-point scale (20*.60)/7=1.71. To the extent these metrics can be compared in light of the aforementioned scaling issues, this would indicate a shift in attitudes that is substantially larger than that which occurred between 1994 and 2014.

5.5 Using Recursive Partitioning to Detect Causal Heterogeneity

We conducted additional analyses to detect possible causal heterogeneity using Athey and Imben's (2016) machine learning approach that employs recursive partitioning. Below we report the result of the LASSO model with change in liberal/conservative ideology scale between the pre- and post-treatment waves as the outcome.

```
#subset variables
  tab = apply(twitter_data[, c(6:20, 59:104)], 2, table)

  for (i in 1:length(tab)){
    print(paste(i, names(tab)[i]))
    print(tab[[i]])
  }

  names(tab)[which(sapply(tab, function(x) length(x) > 10))]
  names(tab)[which(sapply(tab, function(x) length(x) == 1))]
  names(tab)[which(sapply(tab, function(x) length(x) == 0))]

  var_con <- c("birth_year", "followers_count_wave_1", "statuses_count_wave_1",
               "friends_count_wave_1", "family_income")

  var_nom <- c(names(tab)[-c(which(sapply(tab, function(x) length(x) > 10)),
                   which(sapply(tab, function(x) length(x) <= 1)))],
               "state", "religion", "protestant_church", "naics_industry_code")

  #address missing values
  apply(twitter_data[, var_con], 2, function(x) sum(is.na(x)))

  count_miss <- apply(twitter_data[, var_nom], 2, function(x) sum(is.na(x)))
  count_miss[which(count_miss > 0)]

  var_nom <- var_nom[which(!var_nom %in% c("protestant_church"))]

  x_con <- twitter_data[, var_con]
```

```
  x_nom <- data.frame(apply(twitter_data[, var_nom], 2,
                      function(x) factor(as.character(x),
                    levels = names(table(x, useNA = "ifany")))))
  x <- as.data.frame(model.matrix(~.+0, data = as.data.frame(cbind(x_nom, x_con))))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -2.10000 -0.30000  0.00000 -0.00834  0.20000  2.20000      141
```

```
## [1] 956
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
coef(cvfit.lasso, s = "lambda.min")
```

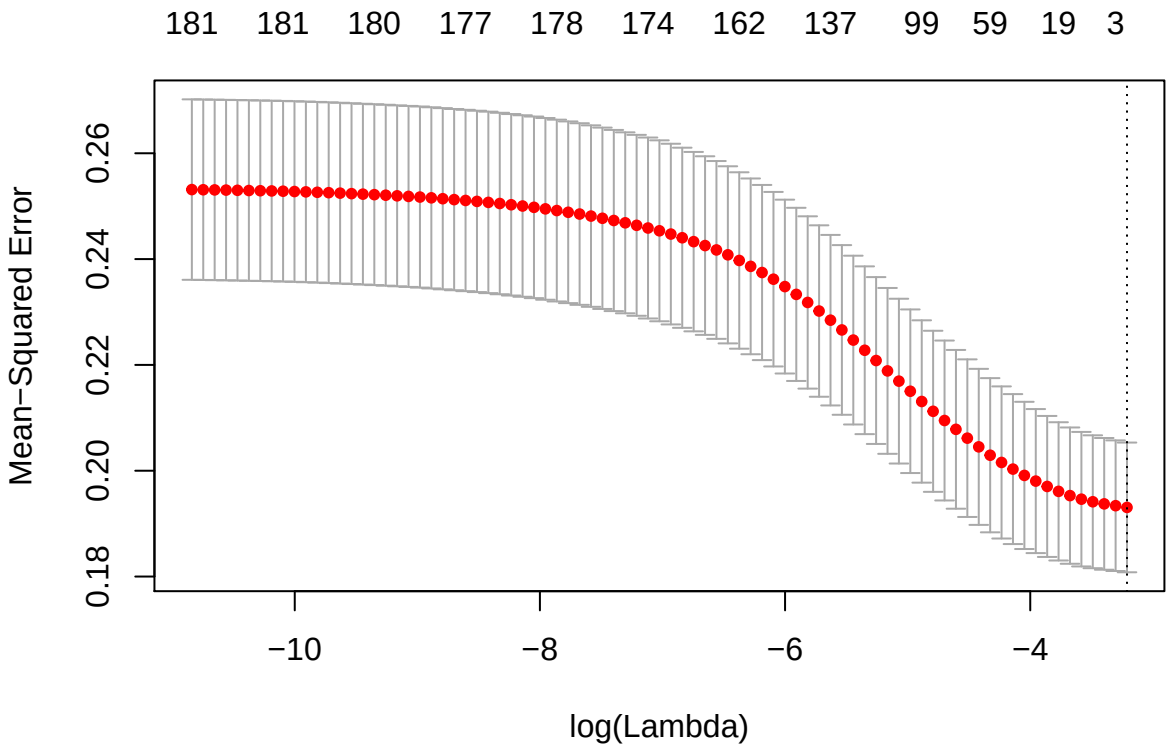
```
## 190 x 1 sparse Matrix of class "dgCMatrix"
##                                                                1
## (Intercept)                                        -0.008338729
## newsint_wave_11                                     .
## newsint_wave_12                                     .
## newsint_wave_13                                     .
## newsint_wave_14                                     .
## newsint_wave_17                                     .
## news_source_newspaper_hard_copy_wave_12             .
## news_source_newspapers_website_wave_12              .
## news_source_news_website_not_newspaper_wave_12      .
## news_source_news_app_mobile_device_wave_12          .
## news_source_email_newsletters_RSS__wave_12          .
## news_source_social.network_websites_wave_12         .
## news_source_blogs_not_major_media_wave_12           .
## news_source_television_wave_12                      .
## news_source_radio_wave_12                           .
## news_source_magazines_wave_12                       .
## news_source_podcasts_wave_12                        .
## news_source_other_wave_12                           .
## news_source_none_of_the_above_wave_12               .
## news_source_dont_know_wave_12                       .
## gender2                                             .
## race2                                               .
## race3                                               .
## race4                                               .
## race5                                               .
## race6                                               .
## race7                                               .
## race8                                               .
## hispanic2                                           .
## multrace_white2                                     .
## multrace_black2                                     .
## multrace_hispanic2                                  .
## multrace_asian2                                     .
## multrace_native_american2                           .
## multrace_middle_eastern2                            .
## multrace_dont_know2                                 .
```

```
## education2                                    .
## education3                                    .
## education4                                    .
## education5                                    .
## education6                                    .
## marital_status2                               .
## marital_status3                               .
## marital_status4                               .
## marital_status5                               .
## marital_status6                               .
## has_children_under_18 2                       .
## speaks_panish 2                               .
## speaks_panish 3                               .
## speaks_panish 4                               .
## employed2                                     .
## employed3                                     .
## employed4                                     .
## employed5                                     .
## employed6                                     .
## employed7                                     .
## employed8                                     .
## employed9                                     .
## employment_otherSelf Employed                 .
## industrynaicsotherAutomotive Manufacturing    .
## industrynaicsotherinformation techology       .
## industrynaicsotherPublic Affairs              .
## industrynaicsotherRetail                      .
## industrynaicsotherWriting                     .
## voter_registration_status2                    .
## voter_registration_status3                    .
## presidential_vote_2016 2                      .
## presidential_vote_2016 3                      .
## presidential_vote_2016 4                      .
## presidential_vote_2016 5                      .
## presidential_vote_2016 6                      .
## presidential_vote_2016 7                      .
## ideology2                                     .
## ideology3                                     .
## ideology4                                     .
## ideology5                                     .
## ideology6                                     .
## born_again2                                   .
## important_religion2                           .
## important_religion3                           .
## important_religion4                           .
## church_attendance2                            .
## church_attendance3                            .
## church_attendance4                            .
## church_attendance5                            .
## church_attendance6                            .
## church_attendance7                            .
## frequency_of_prayer2                          .
## frequency_of_prayer3                          .
## frequency_of_prayer4                          .
```

```
## frequency_of_prayer5                    .
## frequency_of_prayer6                    .
## frequency_of_prayer7                    .
## frequency_of_prayer8                    .
## religpew_t__NA__                        .
## religpew_tBaptist                       .
## religpew_tChristian                     .
## religpew_tEpiscopalian                  .
## religpew_tMethodist                     .
## religpew_tQuaker                        .
## religpew_protestant_tNONE               .
## republican_wave_11                      .
## state 4                                 .
## state 5                                 .
## state 6                                 .
## state 8                                 .
## state 9                                 .
## state10                                 .
## state11                                 .
## state12                                 .
## state13                                 .
## state15                                 .
## state16                                 .
## state17                                 .
## state18                                 .
## state19                                 .
## state20                                 .
## state21                                 .
## state22                                 .
## state23                                 .
## state24                                 .
## state25                                 .
## state26                                 .
## state27                                 .
## state28                                 .
## state29                                 .
## state30                                 .
## state31                                 .
## state32                                 .
## state33                                 .
## state34                                 .
## state35                                 .
## state36                                 .
## state37                                 .
## state38                                 .
## state39                                 .
## state40                                 .
## state41                                 .
## state42                                 .
## state44                                 .
## state45                                 .
## state46                                 .
## state47                                 .
## state48                                 .
```

```
## state49                                 .
## state50                                 .
## state51                                 .
## state53                                 .
## state54                                 .
## state55                                 .
## state56                                 .
## religion 2                              .
## religion 3                              .
## religion 4                              .
## religion 5                              .
## religion 6                              .
## religion 7                              .
## religion 8                              .
## religion 9                              .
## religion10                              .
## religion11                              .
## religion12                              .
## naics_industry_code 2                   .
## naics_industry_code 3                   .
## naics_industry_code 4                   .
## naics_industry_code 5                   .
## naics_industry_code 6                   .
## naics_industry_code 7                   .
## naics_industry_code 8                   .
## naics_industry_code 9                   .
## naics_industry_code10                   .
## naics_industry_code11                   .
## naics_industry_code12                   .
## naics_industry_code13                   .
## naics_industry_code14                   .
## naics_industry_code15                   .
## naics_industry_code16                   .
## naics_industry_code17                   .
## naics_industry_code18                   .
## naics_industry_code19                   .
## naics_industry_code20                   .
## naics_industry_code21                   .
## naics_industry_code22                   .
## naics_industry_code23                   .
## naics_industry_code99                   .
## birth_year                              .
## followers_count_wave_1                  .
## statuses_count_wave_1                   .
## friends_count_wave_1                    .
## family_income                           .
```

```
plot(cvfit.lasso)
```



6 Additional Robustness Checks

6.1 Attrition Bias

The table below indicates there is no evidence of attrition bias by treatment condition.

Table 10: Attrition by Condition

Condition	Pre-Treatment Survey	Post-Treatment Survey
Control	495	413
Treatment	744	656

The table below describes the demographic characteristics of respondents in the pre- and post-treatment surveys.

Table 11: Attrition by Party ID and Demographics

Variable	Pre-Treatment Survey	Post-Treatment Survey
% Republican	43%	42.1%
% Female	51.9%	51.8%
Age (mean)	50.5	51.6
% Northeast	20.3%	21.0%
% South	39.5%	39.2%
% North Central	20.0%	20.4%
% West	21.4%	21.0%

In order to further examine attrition bias created by any of the covariates in our model, we employed the pooling test created by Becketti, Gould, Lillard, and Welch (1988). First, we regressed our outcome variable on the control variables employed in our main analyses alongside a binary indicator variable that describes whether the respondent completed the post-treatment survey. Next, we ran an identical model with interaction terms between the attrition indicator and each of the other variables in the model. An F-test indicates there is no significant difference between the models for Democrats (F=0.85, p<.60) or Republicans (F=1.44, p<.13), suggesting there is no evidence of attrition bias according to these covariates. The code below was used to perform this analysis:

```
#create missing value indicator for post-treatment survey
final_data$wave_5_missing<-0
final_data$wave_5_missing[is.na(final_data$endtime_wave_5)]<-1
table(final_data$wave_5_missing)

#analyze Republican experiment
republicans<-final_data[final_data$party_id_wave_1==2,]

reduced_model<-lm(substantive_ideology_scale_wave_1~
                  wave_5_missing+
                  percent_co_party+
                  political_wave_1+
                  freq_twitter_wave_1+
                  friends_count_wave_1+
                  strong_partisan+
                  birth_year +
                  family_income+
                  education+
                  gender+
                  ideo_homogeneity_offline+
                  northeast+
                  north_central+
                  south,
                  data=republicans)

full_model<-lm(substantive_ideology_scale_wave_1~
                wave_5_missing+
                percent_co_party+
                political_wave_1+
                freq_twitter_wave_1+
                friends_count_wave_1+
                strong_partisan+
                birth_year +
                family_income+
                education+
                gender+
                ideo_homogeneity_offline+
                northeast+
                north_central+
                south+
                wave_5_missing*percent_co_party+
                wave_5_missing*political_wave_1+
                wave_5_missing*freq_twitter_wave_1+
                wave_5_missing*friends_count_wave_1+
```

```
                wave_5_missing*strong_partisan+
                wave_5_missing*birth_year +
                wave_5_missing*family_income+
                wave_5_missing*education+
                wave_5_missing*gender+
                wave_5_missing*ideo_homogeneity_offline+
                wave_5_missing*northeast+
                wave_5_missing*north_central+
                wave_5_missing*south,
                data=republicans)

anova(reduced_model, full_model)


#analyze Democrat Experiment
democrats<-final_data[final_data$party_id_wave_1==1,]

reduced_model<-lm(substantive_ideology_scale_wave_1~
                  wave_5_missing+
                  percent_co_party+
                  political_wave_1+
                  freq_twitter_wave_1+
                  friends_count_wave_1+
                  strong_partisan+
                  birth_year +
                  family_income+
                  education+
                  gender+
                  ideo_homogeneity_offline+
                  northeast+
                  north_central+
                  south,
                  data=democrats)

full_model<-lm(substantive_ideology_scale_wave_1~
                wave_5_missing+
                percent_co_party+
                political_wave_1+
                freq_twitter_wave_1+
                friends_count_wave_1+
                strong_partisan+
                birth_year +
                family_income+
                education+
                gender+
                ideo_homogeneity_offline+
                northeast+
                north_central+
                south+
                wave_5_missing*percent_co_party+
                wave_5_missing*political_wave_1+
                wave_5_missing*freq_twitter_wave_1+
                wave_5_missing*friends_count_wave_1+
```

```
                wave_5_missing*strong_partisan+
                wave_5_missing*birth_year +
                wave_5_missing*family_income+
                wave_5_missing*education+
                wave_5_missing*gender+
                wave_5_missing*ideo_homogeneity_offline+
                wave_5_missing*northeast+
                wave_5_missing*north_central+
                wave_5_missing*south,
                data=democrats)

anova(reduced_model, full_model)
```

At the suggestion of an anonymous reviewer, we also provide an alternative model of attrition bias where the outcome is a dummy variable that indicates whether respondents completed the final survey and the treatment variable is used to predict this outcome in a model where treatment is interacted with every other variable in the model for both our Republican and Democrat experiment. Joint significance tests did not reveal attrition bias for Republicans (p<.48) or Democrats (p<.12). The code used to create these models is below:

```
reduced_model<-glm(wave_5_missing~
                   treat+
                   percent_co_party+
                   political_wave_1+
                   freq_twitter_wave_1+
                   friends_count_wave_1+
                   strong_partisan+
                   birth_year +
                   family_income+
                   education+
                   gender+
                   ideo_homogeneity_offline+
                   northeast+
                   north_central+
                   south,
                   data=republicans,
                   family="binomial")

full_model<-glm(wave_5_missing~
                treat+
                percent_co_party+
                political_wave_1+
                freq_twitter_wave_1+
                friends_count_wave_1+
                strong_partisan+
                birth_year +
                family_income+
                education+
                gender+
                ideo_homogeneity_offline+
                northeast+
                north_central+
                south+
                treat*percent_co_party+
```

```
                treat*political_wave_1+
                treat*freq_twitter_wave_1+
                treat*friends_count_wave_1+
                treat*strong_partisan+
                treat*birth_year +
                treat*family_income+
                treat*education+
                treat*gender+
                treat*ideo_homogeneity_offline+
                treat*northeast+
                treat*north_central+
                treat*south,
                data=republicans,
                family="binomial")

anova(reduced_model, full_model, test="Chisq")


reduced_model<-glm(wave_5_missing~
                   treat+
                   percent_co_party+
                   political_wave_1+
                   freq_twitter_wave_1+
                   friends_count_wave_1+
                   strong_partisan+
                   birth_year +
                   family_income+
                   education+
                   gender+
                   ideo_homogeneity_offline+
                   northeast+
                   north_central+
                   south,
                   data=democrats,
                   family="binomial")

full_model<-glm(wave_5_missing~
                treat+
                percent_co_party+
                political_wave_1+
                freq_twitter_wave_1+
                friends_count_wave_1+
                strong_partisan+
                birth_year +
                family_income+
                education+
                gender+
                ideo_homogeneity_offline+
                northeast+
                north_central+
                south+
                treat*percent_co_party+
                treat*political_wave_1+
```

```
                treat*freq_twitter_wave_1+
                treat*friends_count_wave_1+
                treat*strong_partisan+
                treat*birth_year +
                treat*family_income+
                treat*education+
                treat*gender+
                treat*ideo_homogeneity_offline+
                treat*northeast+
                treat*north_central+
                treat*south,
                data=democrats,
                family="binomial")

anova(reduced_model, full_model, test="Chisq")
```

6.2 Experiment Effects

As with all field experiments, it is possible that respondents in our study shifted their behavior in response to being part of a study. Though such experiment effects typically lead respondents to change their behavior in line with what they perceive to be the expectations of the researchers, it is possible that Republicans in our study had the opposite reaction. That is, Republicans may have expressed more conservative views because they thought the purpose of the study was to make them more liberal, and responded in reactionary fashion by providing more extreme responses in the opposing direction.

As we mentioned in our description of the treatment delivery above, we took several steps to mitigate the likelihood of such experiment effects. First, we employed open-ended questions in two pilot studies that asked people to guess the purpose of our study. Because several respondents in our pilot studies guessed correctly, we made the following shifts to our research design: 1) we employed an ostensibly unrelated survey design in which the treatment delivery occurred one week after the pre-treatment survey (Broockman and Kalla 2016), and did not include reference to the political content of the initial pre-treatment survey; 2) respondents were only shown pictures of landscapes for the first few days of the study, and tweets from those with opposing ideologies were gradually inserted into their Twitter feed thereafter; 3) we revised our informed consent dialogue to de-emphasize that we were academic researchers, we did not describe ourselves as academic researchers or the names of our universities in the first paragraph of our informed consenet dialogue. Instead, as the informed consent dialogue pictured above shows, we only identified ourselves as "researchers" in the third paragraph, and listed the first author's affiliation and contact information in the last paragraph of the informed consent dialogue. At the same time, all invitations to complete surveys came from the same survey firm, so it is possible that some respondents—particularly those who do not take many YouGov surveys—were able to connect our pre- and post-treatment survey to the invitation to follow our study's bots.

Though we cannot rule out the possibility of experiment effects entirely, we believe they are unlikely for the following reasons. First, according to the YouGov Director of Scientific Research, most of the people in the regular YouGov panel take multiple surveys each week—many of which ask them questions about politics. Presumably, connecting our pre-treatment survey to the invitation to follow the Twitter bot would be rather difficult. Second, if people wished to respond in an expressive manner, remembering how they responded to the initial survey would be difficult after one month—particularly for those who regularly take surveys about political issues online. It seems likely that many of those who wish to respond expressively would choose the highest category of response—strongly agreeing with conservative leaning questions and strongly disagreeing with liberal questions after providing answers of different strength in the first wave—in order to demonstrate their displeasure. None of the respondents in our study provided this pattern of response. Finally, the size of the backfire effect we observed increases with level of compliance—whereas one would expect fully compliant

respondents to be no more likely to respond in an expressive manner than partially compliant respondents.

6.3 Outliers

We examined the robustness of our findings to outliers by using Cook's Distance to identify 21 cases which were four times the mean value of Cook's distance for all observations. These findings show the effect reported in the main text of our manuscript is robust to the exclusion of these outliers.

```
first.stage.1 <- lm(substantive_ideology_scale_wave_5 ~
                      treat +
                      substantive_ideology_scale_wave_1+
                      as.factor(bin_maker)+
                      birth_year +
                      family_income +
                      education +
                      gender +
                      ideo_homogeneity_offline +
                      northeast +
                      north_central +
                      south,
                    data=republicans,
                    na.action=na.exclude)
summary(first.stage.1)

republicans$instrumented.perfcomp <- fitted(first.stage.1)

second.stage.1 <- lm(substantive_ideology_scale_wave_5 ~
                       instrumented.perfcomp +
                       substantive_ideology_scale_wave_1+
                       percent_co_party +
                       political_wave_1 +
                       freq_twitter_wave_1 +
                       friends_count_wave_1+
                       strong_partisan +
                       birth_year +
                       family_income +
                       education +
                       gender +
                       ideo_homogeneity_offline +
                       northeast +
                       north_central +
                       south,
                     data=republicans,
                     na.action=na.exclude)

summary(second.stage.1)

cooksd <- cooks.distance(second.stage.1)
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")
influential <- as.numeric(na.omit(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))]))
length(influential)
# identifies 16 cases above accepted val
```

```
republicans$rownumber <- as.numeric(rownames(republicans))
rep_rmalloutlier <- republicans[ ! republicans$rownumber %in% influential,]

#subset variables for models

rep_rmalloutlier<-rep_rmalloutlier[,c("substantive_ideology_scale_wave_5",
                      "perfect_complier",
                      "treat",
                      "substantive_ideology_scale_wave_1",
                      "bin_maker",
                      "percent_co_party",
                      "friends_count_wave_1",
                      "birth_year",
                      "family_income",
                      "education",
                      "gender",
                      "ideo_homogeneity_offline",
                      "northeast",
                      "north_central",
                      "south")]



rep_rmalloutlier<-rep_rmalloutlier[complete.cases(rep_rmalloutlier),]


library(ivpack)
outliers_removed<-ivreg(substantive_ideology_scale_wave_5 ~
                      perfect_complier+
                      substantive_ideology_scale_wave_1+
                      as.factor(bin_maker)+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year +
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south
                      |
                      treat+
                      substantive_ideology_scale_wave_1+
                      as.factor(bin_maker)+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year+
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
```

```
                      north_central+
                      south,
                      data = rep_rmalloutlier)
      #calculate robust standard errors
      output<-robust.se(outliers_removed)[2,]
      output
```

6.4 Post-estimation Weighting by Age

Readers may recall that the demographic characteristics of respondents in our sample track national averages quite well. Because our sample is slighty older than the general U.S. population (median age of our respondents: 50.48, median age of U.S. over-18 population: 46.37), however, it is possible that backfire effects among older Republican respondents drive our findings. To account for this, we ran CACE models with weights for age using data from the American Community Survey. These analyses produced nearly identical results.

```
census_age<-read.csv("~/Desktop/census age data.csv", stringsAsFactors = FALSE)
weights<-census_age[,c("age_group","both_sexes_percent_of_pop")]
names(weights)<-c("age","percent_pop")

final_data$age_percent<-NA
for (i in 1:nrow(final_data)){
final_data$age_percent[i]<-
  nrow(final_data[final_data$birth_year==final_data$birth_year[i],])/
  nrow(final_data)
print(i)
}

library(dplyr)
final_data$age<-2017-final_data$birth_year
weights$age<-gsub(" years","", weights$age)
weights$age<-as.numeric(weights$age)
for_weight<-left_join(final_data, weights)

for_weight$weight<-for_weight$age_percent/for_weight$percent_pop

#subset republicans

republicans<-for_weight[for_weight$party_id_wave_1==2,]
republicans<-republicans[complete.cases(republicans),]

weighted_model<-ivreg(substantive_ideology_scale_wave_5 ~
                      perfect_complier+
                      substantive_ideology_scale_wave_1+
                      as.factor(bin_maker)+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year +
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
```

```
                        south
                        |
                        treat+
                        substantive_ideology_scale_wave_1+
                        as.factor(bin_maker)+
                        percent_co_party+
                        friends_count_wave_1+
                        birth_year+
                        family_income+
                        education+
                        gender+
                        ideo_homogeneity_offline+
                        northeast+
                        north_central+
                        south,
                        data = republicans,
                        weights=weight)

robust.se(weighted_model)

democrats<-for_weight[for_weight$party_id_wave_1==1,]

weighted_model_dems<-ivreg(substantive_ideology_scale_wave_5 ~
                        perfect_complier+
                        substantive_ideology_scale_wave_1+
                        as.factor(bin_maker)+
                        percent_co_party+
                        friends_count_wave_1+
                        birth_year +
                        family_income+
                        education+
                        gender+
                        ideo_homogeneity_offline+
                        northeast+
                        north_central+
                        south
                        |
                        treat+
                        substantive_ideology_scale_wave_1+
                        as.factor(bin_maker)+
                        percent_co_party+
                        friends_count_wave_1+
                        birth_year+
                        family_income+
                        education+
                        gender+
                        ideo_homogeneity_offline+
                        northeast+
                        north_central+
                        south+
                        weight,
                        data = democrats,
                        weights=weight)
```

```
robust.se(weighted_model_dems)
```

7 Alternative Explanations of Results

To further increase our confidence in the findings reported above, we conducted additional analyses of alternative explanations of our results that we present below.

7.1 Additional Twitter Exposure

Though our multiple measures of treatment compliance enable us to make a more accurate estimate of the effect of exposure to Twitter accounts with opposing ideological views, our use of financial incentives to encourage people to pay attention to Twitter might have also increased their exposure to Twitter more broadly. Such additional exposure could influence respondents' political views– either by increasing the overall amount of political messages they are indirectly exposed to or by increasing their exposure to non-political messages that might distract them from politics. The ideal measure would identify the overall exposure of each respondent in our study to political messages on Twitter. Because this measure is not available, we examined whether compliance with our treatment increased the frequency with which respondents report using Twitter. As the code below shows, we detected no significant evidence of increased Twitter exposure using this strategy (Full Compliers, CACE=.45 t=1.28, p<.19, Partial Compliers: CACE: .18, t=1.29, p<.20).

```
twitter_data$twitter_use_change<-twitter_data$how_often_visit_twitter_wave_1-
    twitter_data$how_often_visit_twitter_wave_5

library(ivpack)
summary(ivreg(twitter_use_change~
                perfect_complier
              |
                treat,
              data=twitter_data))

summary(ivreg(twitter_use_change~
                half_complier
              |
                treat,
              data=twitter_data))
```

7.2 Partisan Learning

Related to the issue of additional exposure to Twitter is the possibility that our findings were driven by a process of partisan learning. In other words, it may be that exposure to those with opposing political views helped respondents with little or no interest in politics to learn about their own party's position on the social policies described in our survey.

We took several steps to examine this alternative explanation of our results. First, we examined shifts in the number of people in our treatment condition who responded "Don't know" to any of the ten questions within our liberal/conservative scale between the pre- and post-treatment survey. If partisan learning occurred, we reasoned, fewer people would respond with "Don't know" answers in the post-treatment survey. We did not find evidence of such a trend. Whereas 34 respondents answered "Don't know" to one of the ten questions in

the pre-treatment survey, 37 respondents answered "Don't know" in the post-treatment survey. Table 12 shows the number of respondents with any "Don't know" responses separately for each party-experimental condition combination in the pre-treatment and the post-treatment survey.

Table 12: Prevalence of "Don't Know" Responses in Survey

Party ID	Group	Pre-Treatment	Post-Treatment
Democrats	Treatment	11 (2.96%)	13 (3.50%)
Democrats	Control	5 (2.06%)	5 (2.06%)
Republicans	Treatment	11 (4.01)	13 (4.74%)
Republicans	Control	7 (4.24%)	6 (3.64%)

In addition to this overall trend, within-individual changes do not provide evidence for partisan learning either. While 17 participants in the treatment condition went from responding "Don't know" for any of the social and political issue questions used to create our liberal/conservative scale in the pre-treatment survey to providing a non-missing answer for all questions in the post-treatment survey, 21 exhibited the inverse pattern—providing an answer to all substantive questions pre-treatment but responding "Don't know" to at least one question in the final survey round. Bivariate and multivariate regressions predicting partisan learning operationalized as going from at least one "Don't know" answer to another type of response—or as decreasing the number of "Don't know" responses—similarly suggest no relationship between treatment assignment and partisan learning. Both with and without the controls from our main model, the effect of being assigned to the treatment condition is marginal and not statistically significant. To the extent that partisan learning is defined as increased ability to take a partisan position on political issues, we believe these patterns do not support this mechanism as an explanation of our findings.

Next, we conducted a thorough review of the literature on partisan learning to identify which types of subgroups within our data should be most likely to exhibit partisan learning. Among the most consistent findings in this literature is that those with extreme views are less likely to engage in partisan learning (see, for example, Palfrey and Poole 1987, Delli Carpini and Keeter 1996, and Martin and Desmond 2010). Yet as we show in one of the following sections, we observed nearly identical treatment effects among Republicans with extreme and non-extreme views. Second, there is considerable evidence that Republicans have more knowledge about their party's positions on policy issues than Democrats—in large part because Republicans have taken a more consistent set of ideological positions on a range of issues over time (Grossmann and Hopkins 2016). If this is true, we should have observed increases in liberal attitudes among Democrats in our treatment condition that were larger than the increases in conservative attitudes among Republicans in our treatment condition. Instead, we observed no significant increase in liberalism among Democrats, and a sizeable increase in conservative attitudes among Republicans who were in our treatment condition.

Finally, we examined whether partisan learning might be driving our results via further analysis of the content of the messages retweeted by our bots. More specifically, we reasoned that partisan learning would require cues about the partisan identification of those accounts retweeted by our bots. To assess this possibility, we examined features of the messages retweeted by our bots as well as the Twitter accounts from which they originated. As Table 13 below shows, we found that terms that describe partisan identities are very seldom employed in both the bios of Twitter accounts retweeted by the study's bots as well as the full text of messages. Instead, the opinion leaders retweeted by our bot (elected officials, journalists, media organizations, pundits, and non-profit organizations) tend to eschew explicit partisan identification or labels—perhaps because elected officials do not wish to cue partisanship in order to underscore that they represent all individuals from their geographic region, regardless of party. Similarly, journalists presumably avoid partisan identifiers in order to demonstrate their objectivity in reporting about current events. For an overview of the type of language and words retweeted by our bots, please refer to the tables we provided above.

Table 13: Frequency of Partisan Terms Retweeted by Study's Bots

Partisan Term	Liberal Bios	Cons. Bios	Liberal Tweets	Cons. Tweets

Partisan Term	Liberal Bios	Cons. Bios	Liberal Tweets	Cons. Tweets
"Democrat"	17 (.031%)	0 (0%)	6 (.009%)	4 (.006%)
"Republican"	2 (.003%)*	19 (.042%)	9 (.013%)	7 (.011%)
"Liberal"	0 (0%)	3 (.007%)*	0 (0%)	5 (.008%)
"Conservative"	0 (0%)	42 (.093%)	1 (.001%)	7 (.011%)
"Progressive"	4 (.007%)	1 (.002%)*	1 (.001%)	2 (.003%)
"Dems"	0 (0%)	0 (0%)	16 (.025%)	7 (.011%)
"GOP"	3 (.005%)*	21 (.046%)	18 (.028%)	21 (.033%)

7.3 Extremist Effects

Another alternative explanation of the findings about Republicans reported in the main text of our article is that the effects were driven by extremists. In order to assess this possibility, we ran separate models designed to estimate treatment effects among those Republicans who do not describe themselves as "extremely conservative" on the conventional seven-point liberal/conservative scale question employed by the American National Election Study and many other studies. This question is as follows: "Here is a seven-point scale on which the political views that people might hold are arranged from extremely liberal to extremely conservative. Where would you place yourself on this scale? [Extremely liberal/Liberal/Slightly liberal/Moderate; middle of the road/Slightly conservative/Conservative/Extremely conservative."

These models produced estimates of our treatment effect that are very similar to those reported in the main text of our article. The intent-to-treat effect for non-extremist Republicans was .13 (t=2.49, p<.01). The treatment effect for minimally compliant non-extremist Republicans was .22 (t=2.51, p<.01). The treatment effect for partially-compliant non-extremists Republicans was .33 (t=2.52, p<.01). Finally, the treatment effect for fully-compliant non-extremist Republicans was .62 (t=2.34, p<.02).

```
republicans<-twitter_data[twitter_data$party_id_wave_1==2,]

no_extremists<-republicans[republicans$ideology_seven_point_wave_1<7,]

no_extremists_ITT<-lm(substantive_ideology_scale_wave_5~
                      #treatment assignment variable
                      treat+
                      #pre-treatment ideology score
                      substantive_ideology_scale_wave_1+
                      #% of people followed on Twitter from same party
                      percent_co_party+
                      #% of people in offline networks from same party
                      ideo_homogeneity_offline+
                      #total number of people followed pre-treatment
                      friends_count_wave_1+
                      #demographics
                      birth_year +
                      family_income+
                      education+
                      gender+
                      northeast+
                      north_central+
                      south+
                      #factor variable used to create treatment blocks
```

```
                      as.factor(bin_maker),
                      data=no_extremists)

#calculate robust standard errors
library(lmtest)
library(sandwich)

coefficients<-as.data.frame(coeftest(no_extremists_ITT,
                                     vcov = vcovHC(no_extremists_ITT, type="HC1"))[2:13,1:4])


#Now calculate CACE for varying levels of compliance
library(ivpack)
no_extremists_mc<-ivreg(substantive_ideology_scale_wave_5 ~
                      bot_followers+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year +
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker)
                      |
                      treat+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year+
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker),
                      data = no_extremists)
      #calculate robust standard errors
      robust.se(no_extremists_mc)[2,]



no_extremists_hc<-ivreg(substantive_ideology_scale_wave_5 ~
                      half_complier+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year +
```

```
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker)
                      |
                      treat+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year+
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker),
                      data = no_extremists)
     #calculate robust standard errors
     robust.se(no_extremists_hc)[2,]


no_extremists_fc<-ivreg(substantive_ideology_scale_wave_5 ~
                      perfect_complier+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year +
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
                      north_central+
                      south+
                      as.factor(bin_maker)
                      |
                      treat+
                      substantive_ideology_scale_wave_1+
                      percent_co_party+
                      friends_count_wave_1+
                      birth_year+
                      family_income+
                      education+
                      gender+
                      ideo_homogeneity_offline+
                      northeast+
```

```
                    north_central+
                    south+
                    as.factor(bin_maker),
                    data = no_extremists)
#calculate robust standard errors
robust.se(no_extremists_fc)[2,]
```

7.4 Heterogeneity in Ideological Range of Treatment

As we discuss above, out study's bots identified liberal and conservative opinion leaders using a network-based sampling technique that assumes that those who follow each other on Twitter have similar ideological leanings. Our treatment thus represents "naturally occuring" clusters of liberals and conservatives, rather than a sampling of messages from a uniform distribution of tweets that range from liberal to conservative. The literature on asymmetric polarization, however, suggests that conservatives may espouse more extreme versions of their views than liberals (Grossmann and Hopkins 2016). As an anonymous reviewer pointed out, it is thus possible that our liberal respondents were exposed to a more extreme version of opposing political views than our conservative respondents.

In order to examine this possibility, we took advantage of the fact that we collected both survey data and text data from Twitter. We subsetted respondents who describe themselves as either "extremely liberal" or "extremely conservative" on the conventional seven-point ideology scale employed by the American National Election Study. We then collected all tweets by these individuals, removed high frequency "stop words" such as "and" or "the," and identified the top 500 words that appeared in their tweets.

Next, we compared the list of the top 500 words used by extreme liberals and extreme conservatives to those that appeared in the messages retweeted by our bots during the study period. This analysis revealed that 52.2 percent of the top words used by our respondents who described themselves as extremely conservative were among the top words retweeted by our conservative bot. At the same time, 58.8 percent of the top words used by our respondents who described themselves as extremely liberal were among the top words retweeted by our liberal bot. This analysis indicates that our Democratic respondents were not exposed to a more extreme form of opposing views than our Republican respondents. If anything, this analysis reveals the opposite is true—though further analysis of the content of these tweets revealed that one of the principal reasons for the small discrepancy between the two accounts is that extreme conservatives use fewer political words overall than extreme liberals.

7.5 Gender Effects

Thanks to an anonymous reviewer, we realized that gender may have influenced our findings. Because there are more female opinion leaders who are liberal than conservative—and because men constitute a larger share of Republican respondents than Democratic respondents—the increased conservatism exhibited by some of our Republican respondents may reflect their beliefs that women should not occupy positions of power. Though such beliefs may only represent a small minority of Republicans, it is noteworthy that there are sixty-two female members of the U.S. House of Representatives who are Democrats and twenty-two who identify as Republicans. We also note that derogatory comments related to Hillary Clinton's gender were not uncommon among conservative opinion leaders both before and after the 2016 presidential election.

In order to examine this alternative—though complimentary—explanation of our findings, we passed the names associated with all messages produced by our liberal and conservative bots through the Genderize Application Programming Interface. This tool compares given names to databases that describe the number of men and women who have such names in broader populations in order to make predictions about the likely gender of those who have a particular name. Our analysis revealed that 35.3% of the messages retweeted by the liberal bot were produced by those whose gender was predicted to be female by the Genderize API compared to 26.3% of the messages produced by the conservative bot. Because the difference between the

two bots is highly significant ($p<.001$), we cannot rule out the possibility that some of the backfire effect we observed was part of negative response to women in positions of power– particularly because we did not measure attitudes about gender among respondents in our study. We note, however, that the magnitude of the gender difference just described is considerably less than one might expect by comparing the gender composition of the U.S. House of Representatives. This likely reflects the substantial numbers of female opinion leaders in non-elected positions such as journalism, advocacy organizations, think tanks, or other civil society groups that were retweeted by our bot.

It is also possible that race effects contribute to our findings. Unfortunately, machine learning algorithms to measure race are not reliable and we are therefore unable to conduct a careful analysis of this alternative explanation here.

7.6 Age and Social Media Usage

Because our sample is slightly older than the general U.S. population, an anonymous reviewer recommended that we examine whether older people use social media sites in a different manner than younger users. In order to assess this issue, we obtained data from the Pew Research Center's "American Trends Panel (Wave 19)" which was conducted from July-August 2016. This report is available at this link

First, we examined responses to the following question by age: "How often do you get political news on social media?" As Table 14 below shows, the Pew data indicate no significant differences across age groups (chi-square with six degrees of freedom = 5.993, $p = 0.424$).

Table 14: Consumpton of Political News on Social Media by Age Group

Age	Often	Sometimes	Hardly ever/Never	Total
18-29	50.93%	34.78%	14.29%	100%
30-49	43.8%	37.78%	19.04%	100 %
50-64	50.98%	30.88%	18.14%	100%
65+	44.44%	33.33%	22.23%	100%

Next, we examined whether age shapes the strength of social media echo chambers. Table 15 below describes the results of the following Pew survey question by age group: "Do most of the people you follow on Twitter have have Similar political beliefs to you/Different political beliefs from you/A mix of political beliefs/No Answer." Once again, we discovered no statistically significant differences by age (chi-square with nine degrees of freedom = 13.733, $p = 0.132$).

Table 15: Repondent Estimates of Twitter Echo Chamber Strength by Age

Age	Similar Beliefs	Different Beliefs	Mix of Beliefs	No Answer	Total
18-29	24.22%	5.59%	32.92%	37.27%	100%
30-49	19.14%	2.86%	45.71%	32.29%	100 %
50-64	17%	2%	46%	35%	100%
65+	21.43%	1.43%	47.14%	30%	100%

Finally, we examined data from the same Pew survey on respondents' experience of out-partisan contact on social media. Table 16 below reports responses to the following question by age group: "In your experience, when you talk about politics with people on social media who you disagree with, do you generally find it to be ... Interesting and informative/stressful and frustrating/no answer." Once again, we found no significant differences across age groups (chi-square with three degrees of freedom = 0.625, $p = 0.891$).

Table 16: Assessments of Out-Partisan Contact on Social Media Sites by Age

Age	Interesting and Informative	Stressful and Frustrating	Total
18-29	35.03%	64.97%	100%
30-49	33.82%	66.18%	100%
50-64	31.41%	68.59%	100%
65+	31.82%	68.18%	100%

8 Additional Outcome Measures

As we mentioned in the main text of our manuscript and in our pre-registration statement, the ten-item ideological consistency scale that we analyze throughout this article was but one of the political attitudes that we measured in our study. In addition to this outcome, we collected multiple measures of affective polarization, expressive partisanship, and bipartisan engagement.

Our measures of affective polarization include conventional thermometer and social distance ratings for the opposing political parties as well as a set of variables that ask respondents whether members of the opposing political party are a) patriotic, b) intelligent, c) honest, d) open-minded, d) generous, e) close-minded, f) hypocritical; and g) selfish or mean. We measured expressive partisanship using the following questions: a) "How important is being a Democrat/Republican to you?"; b) How well does the term (Democrat/Republican) describe you?; c) When talking about (Democrats/Republicans) how often do you use the term "we" instead of "they"? To measure desire for bipartisan engagement we asked respondents the following questions: a) "It is important to talk about political issues with (Republicans/Democrats)"; and, b) I make efforts to watch TV shows, or listen to radio that (Republicans/Democrats) usually consume in order to better understand (Republicans/Democrats) views. Finally, we asked the following question to measure overall awareness of the opposing political party: "Thinking about the United States as a whole, how many Americans do you think DISAGREE with your political views?"

We plan to report the effects of our treatment on these variables in future publications. Though some of our scales proved unreliable, our preliminary analyses indicate there is substantial variation in treatment effects across these outcomes.

References

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27). National Acad Sciences: 7353–60.

Bafumi, Joseph, and Michael C Herron. 2010. "Leapfrog Representation and Extremism: A Study of American Voters and Their Members in Congress." *American Political Science Review* 104 (3). Cambridge University Press: 519–42.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91. doi:10.1093/pan/mpu011.

Broockman, David, and Joshua Kalla. 2016. "Durably reducing transphobia: A field experiment on door-to-door canvassing." *Science (New York, N.Y.)* 352 (6282). American Association for the Advancement of Science: 220–4. doi:10.1126/science.aad9713.

Dimock, Michael, and Doherty Carroll. 2014. "Political Polarization in the American Public: How Increasing Ideological Uniformity and Partisan Antipathy Affect Politics, Compromise, and Everyday Life." *Pew Research Center Report*.

Fiorina, Morris P, Samuel J Abrams, and Jeremy Pope. 2006. *Culture War?: The Myth of a Polarized America*. Longman Publishing Group.

Gentzkow, Matthew, and Jesse M Shapiro. 2011. "Ideological Segregation Online and Offline." *The Quarterly*

Journal of Economics 126 (4). MIT Press: 1799–1839.

Grossmann, Matt, and David A Hopkins. 2016. *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*. Oxford University Press.

Grönlund, Kimmo, Kaisa Herne, and Maija Setälä. 2015. "Does Enclave Deliberation Polarize Opinions?" *Political Behavior* 37 (4). Springer: 995–1020.

Gupta, Pankaj, Ashish Goel, Jimmy Lin, Aneesh Sharma, Dong Wang, and Reza Zadeh. 2013. "Wtf: The Who to Follow Service at Twitter." In *Proceedings of the 22nd International Conference on World Wide Web*, 505–14. ACM.

Hill, Seth J, and Chris Tausanovitch. 2015. "A Disconnect in Representation? Comparison of Trends in Congressional and Public Polarization." *The Journal of Politics* 77 (4). University of Chicago Press Chicago, IL: 1058–75.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jessee, Stephen A. 2012. *Ideology and Spatial Voting in American Elections*. Cambridge University Press.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176). American Association for the Advancement of Science: 1203–5.

Luskin, Robert C, James S Fishkin, and Kyu S Hahn. 2007. "Deliberation and Net Attitude Change." In *ECPR General Conference, Pisa, Italy*, 6–8.

Tausanovitch, Chris, and Christopher Warshaw. 2013. "Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities." *The Journal of Politics* 75 (2). Cambridge University Press New York, USA: 330–42.