Replicating # No2Sectarianism

Professor: Adeline Lo // TA: Dillon Laaker
Group Project by Jess Esplin, J. Clinton Rooker, & Ethan vanderWilden

Presented in PS812 on 12/10/2020

ABSTRACT

We replicate Siegal and Badaan's (2020) paper in which they evaluate the efficacy of various types of counter-speech interventions in reducing sectarian hate speech online. We analyze data from their Twitter experiment, replicate their findings, and consider further extensions of the study. Siegal and Badaan conducted an experiment on Arab Twitter using a fake account to reply to users who regularly tweeted sectarian content with messages sanctioning such content and invoking a prime to a superordinate group identity (religious or nationalist), as well as some which referenced elite (religious or political leaders) condemnation of such content. We find support for Siegal and Badaan's (2020) findings - specifically, we find statistical significance confirming their conclusion that elite-endorsed messages priming common religious identity were most consistently effective in reducing sectarian hate speech. We do not find evidence to reject the null hypotheses for the other treatment variations. We explore potential improvements that could be made, verify the robustness of findings, and propose possible extensions to address unanswered questions on this topic.

I. INTRODUCTION & EXPERIMENTAL DESIGN

We[1] replicate the findings of Siegal and Badaan (2020)[2], using their replication materials available on the Harvard Dataverse. They conducted an experiment on Twitter to evaluate various interventions on online Arab sectarian hate speech, where sectarianism is defined as "pro-ingroup bias based on affiliation to a particular confessional or religious group" (p. 839). Specifically, the authors assessed whether counter-speech messages on Twitter priming a common national or religious identity, some with elite endorsements and some without, reduced sectarian hate speech online. They also conducted a survey experiment in Lebanon to evaluate Arab citizen responses to counter-speech interventions, but we address only the Twitter experiment here.

Siegal and Badaan (2020) used a sockpuppet (fake Twitter account that they control) to tweet counter-speech messages at Arab Twitter users who regularly tweet sectarian content. The sockpuppet was designed to appear as an average Sunni male Twitter user from the Arabian Gulf and replied to selected Twitter users with one of five randomly assigned messages. Four of these messages sanctioned the Twitter user and included a prime for an identity: (1) common Muslim religious identity, (2) common Arab national identity, (3) common Muslim religious identity with an endorsement from religious elites, or (4) common national identity with an endorsement from political elites, while (5) the last treatment condition sanctioned without an identity prime. Another group of Twitter users were assigned to a control group and did not receive a message. The differences in pre- and post-treatment sectarian tweets are compared to differences in sectarian tweets over the same period in the control group.

The theoretical motivation of this experiment is primarily based on identity recategorization: the authors argue group identities are malleable and therefore intergroup bias can be effectively reduced through

[1] The authors note that work on this project was split equally; authors are listed alphabetically by last name.

[2] Siegel, Alexandra A., and Vivienne Badaan. 2020. "# No2Sectarianism: Experimental Approaches to Reducing Sectarian Hate Speech Online." American Political Science Review 114:3, 837-855.

this process “by which members of different groups are primed to view themselves as part of a single, more inclusive superordinate group” (Siegal and Badaan 2020). Furthermore, Siegal and Badaan argue the social psychology literature on social norms and ingroup identity suggest identity recategorization “may be especially effective when trusted elites – who have a great deal of power to shape norms or boundaries of group membership – deliver the message” (p. 840). In sum, they posit that priming a superordinate group identity in a message from a trusted elite is likely to be the most effective intervention to counter hate speech.

II. HYPOTHESIS

Siegal and Badaan (2020)’s four hypotheses relevant to the Twitter experiment fall under two categories: (A) priming group identity without elite endorsement and (B) priming group identity with elite endorsement. They present this variety of hypotheses following the same format in which the null hypothesis states there is no effect of the treatment on Twitter user subjects’ tweets containing sectarian hate speech, versus hypotheses in which treatment will decrease the number of tweets from subjects containing sectarian hate speech. Thus, our main hypothesis is as follows: Receiving a response to a hateful sectarian message (that primes a group identity and/or highlights support from elites) will make individuals less likely to tweet hateful sectarian messages in the future – relative to receiving all other primes, a sanctioning message with no prime, or no reply.

- H_0: $\bar{x}_{treatment} - \bar{x}_{control} \geq 0$
- H_A: $\bar{x}_{treatment} - \bar{x}_{control} < 0$

III. DATA

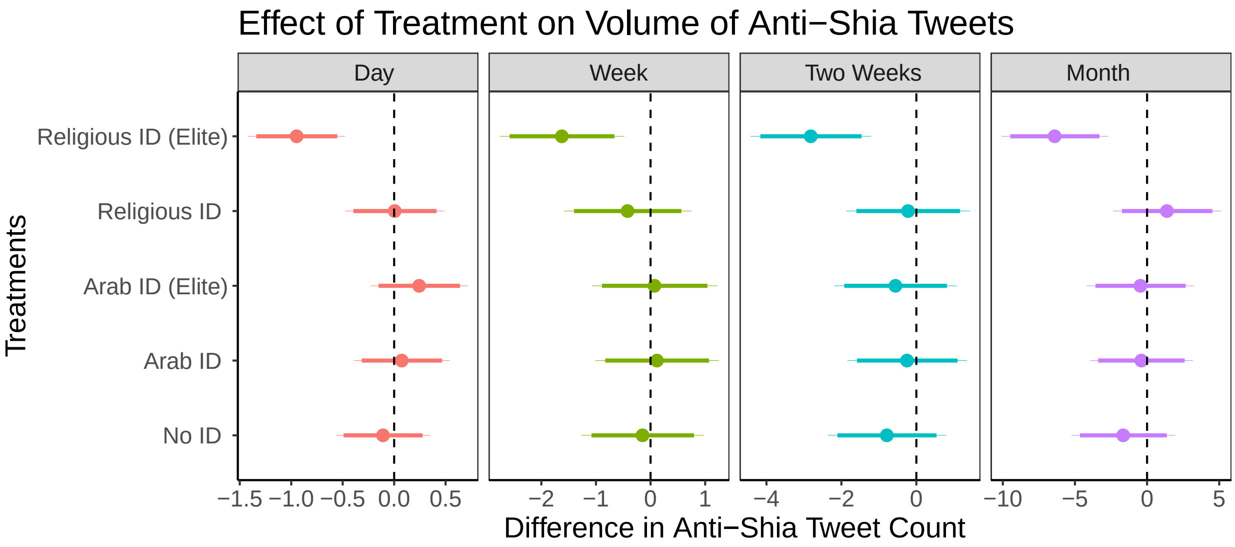
Data for this experiment were collected in the form of tweets from study subjects, comprising 9,957 Twitter users who were identified as Arabs that had recently tweeted anti-Shia slurs. Twitter users with more than 10,000 followers were excluded from the experiment. The primary data are counts of anti-Shia tweets before and after treatment.

IV. RESULTS

Our main results presented above confirm the original authors’ “Elite Common Religious Identity Hypothesis” (p. 841). This states that receiving a message that primes a common religious identity while also highlighting religious elite support will decrease the user’s likelihood of sending sectarian hate tweets in the future.

For each time period presented under this treatment, we find statistically significant differences in post- and pre-treatment hate tweets ($p<0.01$ for all four time intervals). One day after an intervention that primes a common religious message with elite support, our results show an average treatment effect of -0.948 tweets. Similarly, one week after this intervention, we notice an average decrease of 1.625 sectarian hate tweets; two weeks after this intervention, we notice an average decrease of 2.817 sectarian hate tweets; and one month after this intervention, we notice an average decrease of 6.405 sectarian hate tweets. With regard to the other three treatments (the first three counter-speech interventions noted above), we do not have evidence to reject the null hypotheses that post- and pre-treatment sectarian hate tweets are significantly different. As demonstrated in the figure above, confidence intervals for the average treatment effects of these treatments each contain zero.

FIGURE 1. Treatment Effect Confidence Intervals



V. DISCUSSION

One area for improvement in the research design is the refinement of the social norms testing. The authors posit that their tweet intervention would be most effective for users with a low number of anti-Shia friends within their network. To assess the prevalence of anti-Shia tweets within a network, the authors compiled the network data for each of their subjects and cross-validated this list with a dataset of all anti-Shia tweets in the pre-treatment period. Then, for each subject, the authors calculated the number of friends in their network, and measured the treatment effect of their intervention for those subjects with a high and low number of anti-Shia Twitter friends, as indicated by being above or below the median number of anti-Shia friends within the network.

Rather than using the median number of friends as the point of reference for anti-Shia network size, we suggest the authors employ a proportional test to better evaluate the treatment effect within different networks. Unfortunately, the authors employed a somewhat ambiguous and unclear definition of "anti-Shia friends" versus "followers" for their subjects. As such, it is unclear whether a friend within the network is an account followed by the subject, or an account that follows the subject. This makes it difficult to assess what the authors are measuring when they discuss network size and friends within the network. To clarify this test, a proportional metric measuring anti-Shia network saturation could be employed.

In order to better evaluate the social norms mechanism, the authors may consider clarifying their definition of networks and friends, and the relationship between these variables and total followers. This would enable the application of a proportional test of anti-Shia network saturation, which may reveal disparate treatment effects relative to the proportion of one's network being composed of anti-Shia friends. In practice, there is reason to believe that a proportional test is more robust than the median test. For example, if a subject has a total network of 1,000 friends, and 50 of those friends produce anti-Shia content, their network norms may operate differently than a subject with a network of 100 friends, with 50 users generating similar content. As such, a proportional test of relative network saturation would more accurately reflect the salience and prevalence of anti-Shia sentiment within a network.

Additionally, we believe several assumptions put forth by the authors warrant further scrutiny. First, the population of Twitter users within Arab countries may be skewed in such a way that the subjects in this experiment do not represent the broader Arab population. Northwestern's annual Middle East Media Use survey indicates that Twitter users in Arab states skew younger and more educated\footnote{Dennis, E., Martin, J., Lance, E., & Hassan, F. (2019). Media Use in the Middle East: A Seven-Nation Survey. Retrieved from http://www.mideastmedia.org/survey/2019/uploads/file/NUQ_Media_Use_2019_WebVersion_FNL_051219%5B2%5D(1).pdf}. Moreover, men report having more followers on average than women in the media use survey. Adding to the issue of representativeness is the lack of any descriptive demographic statistics on the study's subjects. Admittedly, this is difficult to do, as it would require a researcher to ascribe racial or ethnic characteristics and gender identity. Further complicating matters is validating the veracity of a user's profile picture and the additional difficulty of this in instances where a profile picture is not provided. Those challenges notwithstanding, it is important to understand the level of generalizability from the results, and there may be underlying issues with the population on Twitter. Therefore, the Twitter users examined within the study may be poorly representative of the broader population, which raises questions about the applicability of the treatment effects to the broader population of social media users. Of note, the authors do not provide any socioeconomic or demographic descriptive data regarding their subjects in the Twitter experiment.

A second assumption rests on the length of the treatment tweet. The assumption remains untested throughout the paper and there is reason to doubt its validity. Although the four treatment tweets are roughly the same length, the tweets including elite endorsement are nominally longer than the non-elite treatments. As the elite common religious identity treatment is slightly longer than the common religious and national identity treatments, it may influence the effect of the treatment. Importantly, the elite common religious identity tweet is the only treatment that displays a significant effect, so the assumption that tweet length exerts a uniform effect on the study's subjects may undermine the findings.

We perform two robustness checks. First, we replicate the authors' refined test that looks only at users who have the median or fewer followers. The authors acknowledge that some Twitter users may have too many followers to be affected by the various treatments, thus accounting for the observed null effects. However, examining only those users who have the median or fewer followers, the results presented in Table and Figure 1 hold. Still, only the prime of elite-endorsed common religious identity produces a statistically significant decrease in sectarian hate tweets. The average treatment effects and confidence intervals can be found in our Appendix under Table X and Figure 2.

We also perform an original robustness check that examines treatment for those with networks that are the most saturated with anti-Shia sentiments. In line with the above suggested "proportional test" of anti-Shia friends, we approximate what such a robustness check would look like by using the number of Twitter followers as a proxy for overall friends. Here, we are assuming that followers roughly equal the amount of friends that a user might have. This is clearly not the case, as the number of anti-Shia friends is greater than the number of followers in some cases. Still, however this robustness check gives an initial look into the saturation of anti-Shia sentiment within one's Twitter network.

Following the logic of network type (referring to the presence of social norms about anti-Shia sentiments), we conduct robustness checks that separate network type by network saturation, rather than raw anti-Shia friend counts. The results presented in Tables and Figures 3 and 4 in the appendix largely align with our main findings. Still, the elite-endorsed message that primes common religious identity is the only statistically significant treatment, although in more anti-Shia saturated networks, the effects are slightly dampened.

VI. NEXT STEPS

After replicating the authors' main findings, we identified multiple paths forward for further analysis. One potential avenue for future assessment is the application of a follow-up treatment tweet, sent to users who engage with the initial treatment. For example, the paper explores a one-time intervention, where the research team responds to an anti-Shia tweet with a single treatment response. We were interested in whether this approximates a natural exchange on Twitter, or if there are different ways that individuals would engage in such a situation. It may be worthwhile to observe whether anti-Shia Twitter users respond to the intervention, in which case a subsequent treatment tweet could be sent to the user. This would allow researchers the opportunity to measure the propensity of a Twitter user to engage with the treatment, measure the relative impact of multiple treatments, and create a more realistic Twitter dialogue.

Moreover, we are interested in whether the impact of the treatment is confined solely to Twitter. While the mode of engagement differs across social media platforms, we believe there are ways to approximate similar exchanges in different domains. Future analyses could apply a treatment comment to posts on Facebook or Instagram and examine whether the treatment displays similar effects. In the past several years, there has been a decrease in Twitter usage in Arab countries and a concurrent increase in Facebook and Instagram users in the region. As interactions increase in these digital spaces, researchers should examine how anti-Shia attitudes propagate among users, and interact between users.

Finally, future research should clearly define the relationship between anti-Shia friends within a user's network and assess how treatment effects behave among varying degrees of anti-Shia network saturation. We have proposed a proportional measure, where a user's anti-Shia network saturation is determined by the number of followers and followed accounts that express anti-Shia sentiment divided by the total number of followers and followed accounts. This could allow for a more accurate assessment of anti-Shia attitudes within a user's network than the current median metric that is employed.

Does quality of interaction matter instead of merely a "one and done" comment? Sustained contact, thereby longer tweets? Exploring this phenomenon on other platforms (as usage on Twitter has been declining relative to Facebook and Instagram)?

VII. APPENDIX

1. Main Effects Test

A note on our regression tables: In their paper, the original authors opt to display their results graphically with 90 and 95% confidence intervals. This demonstrates which treatments result in differences in sectarian tweets that are significantly different from zero. Here, we offer additional information as to the exact coefficients (average treatment effects) and standard errors for treatment effects in the form of a regression table. Importantly, the difference of means test is functionally the same as this simple OLS regression with a binary independent variable.

Table 1: Effect of Treatment on Volume of Anti-Shia Tweets

	Difference in Anti-Shia Tweets			
	Day	Week	Two Weeks	One Month
	(1)	(2)	(3)	(4)
Arab ID	0.074	0.116	−0.249	−0.403
	(0.236)	(0.574)	(0.811)	(1.811)
Religious ID	0.006	−0.421	−0.221	1.376
	(0.244)	(0.594)	(0.836)	(1.896)
Arab ID (elite)	0.243	0.072	−0.557	−0.461
	(0.239)	(0.584)	(0.828)	(1.889)
Religious ID (elite)	−0.948***	−1.625***	−2.817***	−6.405***
	(0.238)	(0.580)	(0.816)	(1.867)
No ID	−0.108	−0.147	−0.787	−1.657
	(0.232)	(0.567)	(0.798)	(1.820)
Constant	0.015	0.298	1.304**	1.816
	(0.176)	(0.431)	(0.612)	(1.404)
Observations	952	944	922	795
R^2	0.035	0.015	0.019	0.029
Adjusted R^2	0.030	0.010	0.014	0.023

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

2. Robustness Checks

(a) Median Subset

Logic: It is possible that some subjects are very popular on Twitter - despite the original authors' exclusion criteria of less than 10,000 followers for experimental subjects, a Twitter user with followers on the higher end of this scale is still quite popular. With popularity, we might expect such subjects to be less likely to be impacted by a single counter-sectarian tweet, given that their influence is comparatively large. This could be beneath the majority null result that we found. Here, we subset the data such that only subjects with less than the median number of followers are included in the dataset.

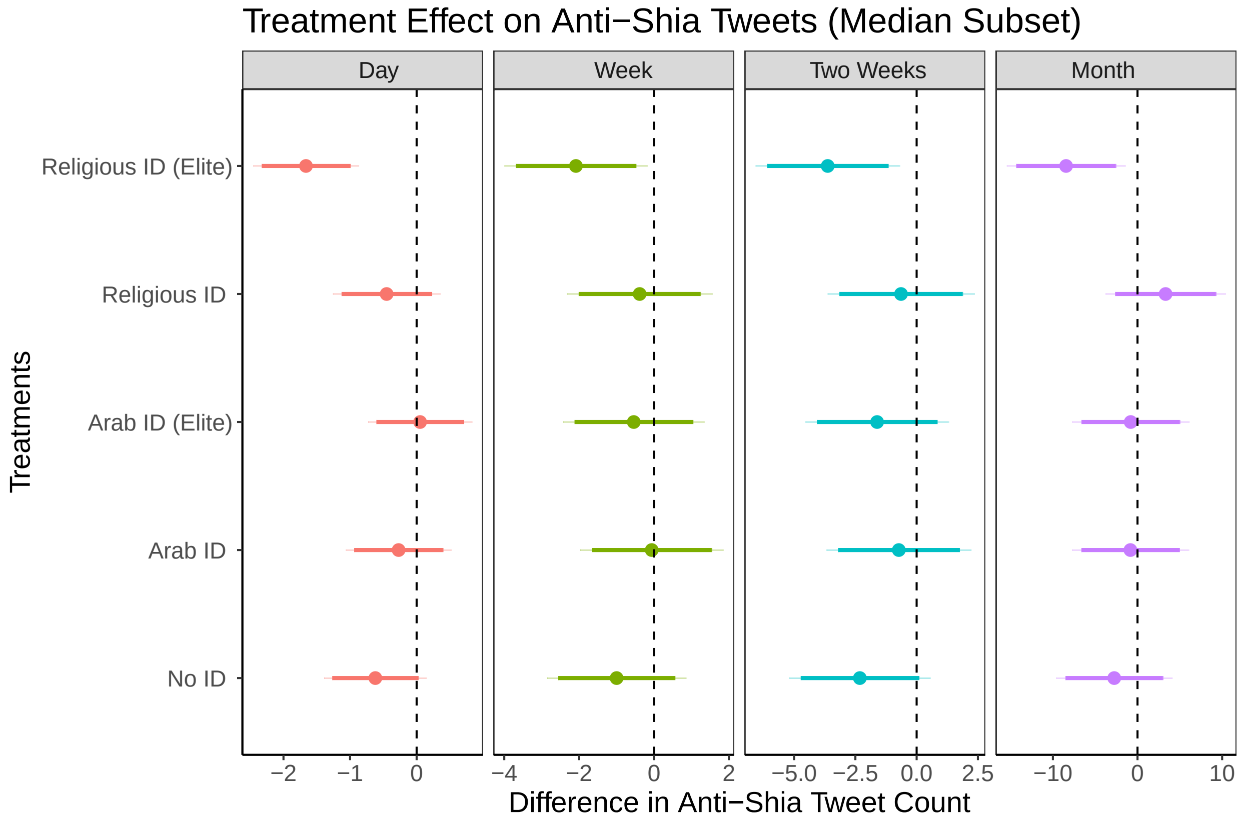


FIGURE 2. Treatment Effect Confidence Intervals Using Median Subset

Table 2: Effect of Treatment on Volume of Anti-Shia Tweets (Median or Fewer Follower Subset)

	Difference in Anti-Shia Tweets			
	One Day	One Week	Two Weeks	One Month
	(1)	(2)	(3)	(4)
Arab ID	−0.271	−0.061	−0.727	−0.838
	(0.405)	(0.974)	(1.504)	(3.518)
Religious ID	−0.450	−0.384	−0.637	3.316
	(0.411)	(0.990)	(1.527)	(3.614)
Arab ID (elite)	0.052	−0.541	−1.614	−0.806
	(0.398)	(0.960)	(1.490)	(3.527)
Religious ID (elite)	−1.662***	−2.086**	−3.630**	−8.441**
	(0.404)	(0.974)	(1.501)	(3.573)
No ID	−0.620	−0.999	−2.318	−2.751
	(0.392)	(0.946)	(1.466)	(3.494)
Constant	0.357	0.691	1.840	2.238
	(0.311)	(0.752)	(1.180)	(2.817)
Observations	477	473	461	403
R^2	0.058	0.016	0.020	0.037
Adjusted R^2	0.048	0.005	0.009	0.025

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

(b) Sectarian Saturation

(i) Low Anti-Shia Saturation

Logic: Susceptibility to the interventions may be derived by the social norms within Twitter networks. If one's Twitter network is overwhelmingly anti-Shia, we might expect the intervention is not as effective - potentially accounting for some of the null results. In this test, we divide the number of anti-Shia friends by the total followers for each subject. We use this as a proxy to measure anti-Shia network saturation. The below test examines the treatment effect among users in low anti-Shia-saturated networks (cutoff point at median saturation).

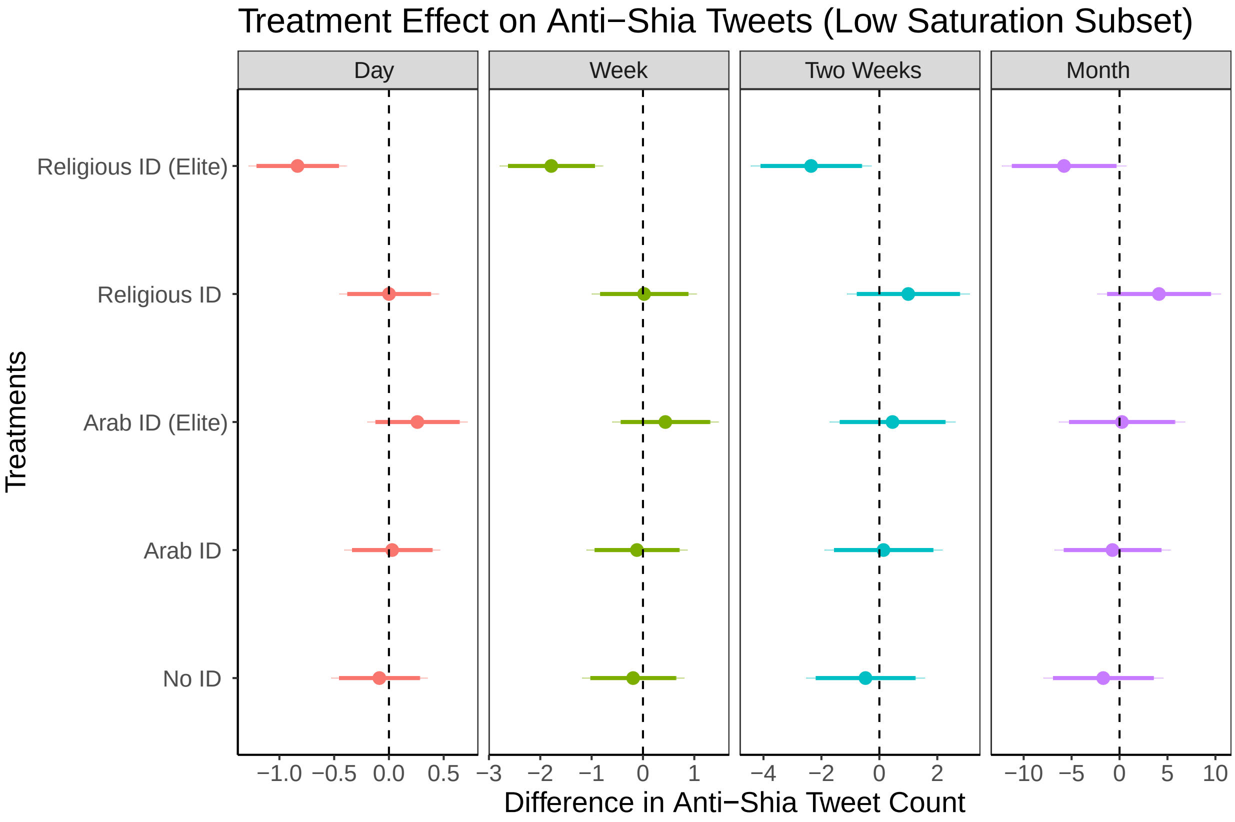


Figure 3. Treatment Effect Confidence Intervals with Low Anti-Shia Saturation

Table 3: Effect of Treatment on Volume of Anti-Shia Tweets (Median or Less Saturated Network Subset)

	Difference in Anti-Shia Tweets			
	One Day	One Week	Two Weeks	One Month
	(1)	(2)	(3)	(4)
Arab ID	0.029	−0.118	0.144	−0.738
	(0.223)	(0.501)	(1.037)	(3.084)
Religious ID	0.001	0.023	0.997	4.109
	(0.231)	(0.520)	(1.078)	(3.279)
Arab ID (elite)	0.260	0.436	0.453	0.248
	(0.233)	(0.527)	(1.106)	(3.352)
Religious ID (elite)	-0.834^{***}	-1.786^{***}	-2.357^{**}	-5.789^{*}
	(0.228)	(0.512)	(1.061)	(3.307)
No ID	−0.087	−0.192	−0.479	−1.700
	(0.224)	(0.506)	(1.043)	(3.181)
Constant	−0.041	0.311	0.729	1.789
	(0.164)	(0.369)	(0.770)	(2.338)
Observations	476	472	457	371
R^2	0.056	0.047	0.027	0.026
Adjusted R^2	0.046	0.037	0.016	0.012

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

ii) High Anti-Shia Saturation

Logic: We now examine results with the high anti-Shia network saturation subjects to better understand whether the treatment results hold.

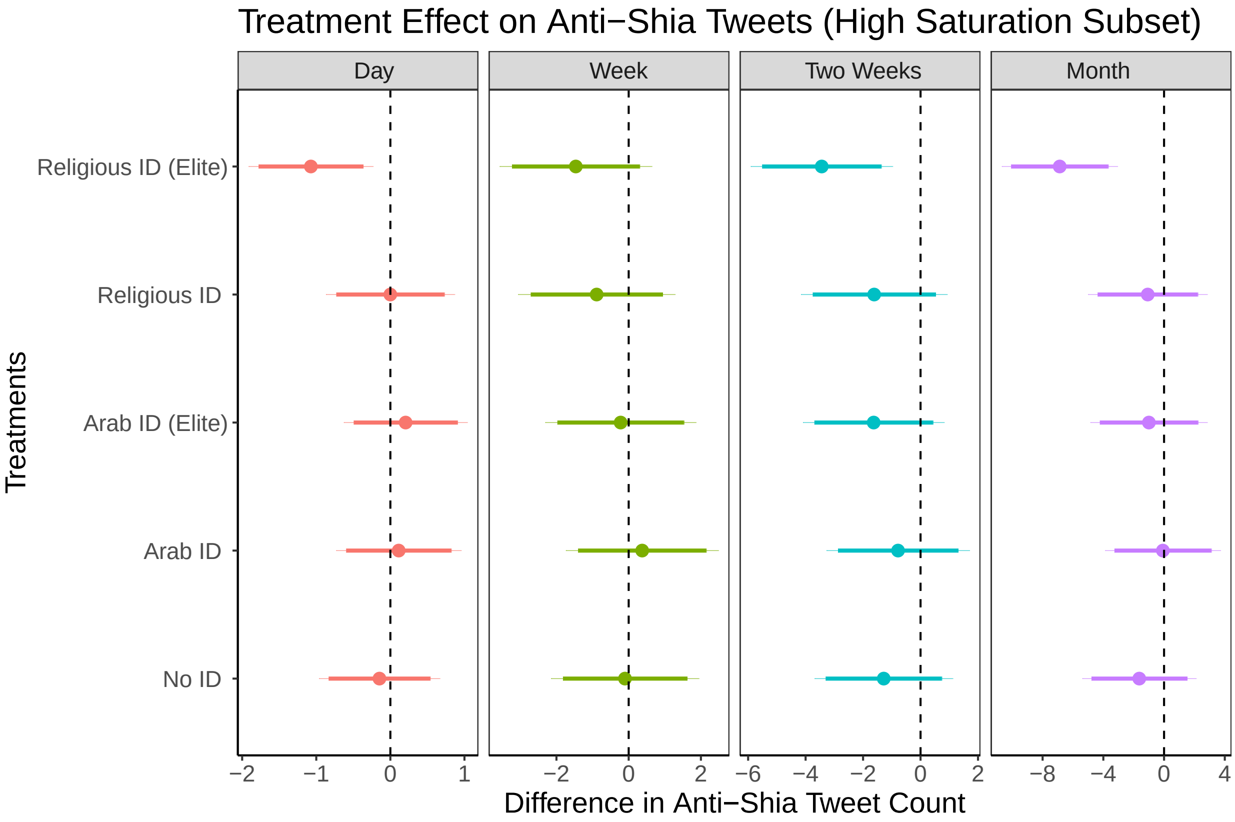


Figure 4. Treatment Effect Confidence Intervals with High Anti-Shia Saturation

Table 4: Effect of Treatment on Volume of Anti-Shia Tweets (Median or Greater Saturated Network Subset)

	Difference in Anti-Shia Tweets			
	One Day	One Week	Two Weeks	One Month
	(1)	(2)	(3)	(4)
Arab ID	0.113	0.374	−0.783	−0.079
	(0.430)	(1.077)	(1.269)	(1.938)
Religious ID	−0.0002	−0.886	−1.614	−1.075
	(0.443)	(1.108)	(1.298)	(2.003)
Arab ID (elite)	0.206	−0.223	−1.632	−0.999
	(0.425)	(1.064)	(1.253)	(1.963)
Religious ID (elite)	−1.073**	−1.464	−3.439***	−6.874***
	(0.428)	(1.074)	(1.259)	(1.943)
No ID	−0.147	−0.102	−1.281	−1.634
	(0.416)	(1.044)	(1.226)	(1.912)
Constant	0.085	0.281	2.036**	1.848
	(0.327)	(0.825)	(0.974)	(1.537)
Observations	476	472	465	424
R^2	0.029	0.010	0.020	0.051
Adjusted R^2	0.019	−0.001	0.009	0.040
Note:			*p<0.1; **p<0.05; ***p<0.01	