Internet Research Agency Twitter activity predicted 2016 U.S. election polls

Damian J. Ruck,[1*] Natalie Manaeva Rice,[2] Joshua Borycz,[2,3]R. Alexander Bentley,[1]

[1]Department of Anthropology, University of Tennessee, Knoxville, TN USA
[2]Center for Information and Communication Studies, University of Tennessee, Knoxville, TN USA
[3]Science and Engineering Library, Vanderbilt University, Nashville, TN, USA

[*]To whom correspondence should be addressed; E-mail: druck@utk.edu.

In 2016, the Internet Research Agency (IRA) deployed thousands of Twitter bots that released hundreds of thousands of English language tweets. It has been hypothesized this affected public opinion during the 2016 U.S. presidential election. Here we test that hypothesis using Vector Auto-regression comparing time series of election opinion polling during 2016 versus numbers of re-tweets or 'likes' of IRA tweets. We find that changes in opinion poll numbers for one of the candidates were consistently preceded by corresponding changes in IRA re-tweet volume, at an optimum interval of one week before. In contrast, the opinion poll numbers did not correlate with future re-tweets or 'likes' of the IRA Tweets. We find that the release of these tweets parallel significant political events of 2016 and that approximately every 25,000 additional IRA re-tweets predicted a 1% increase in election opinion polls for one candidate. As these tweets were part of a larger, multimedia campaign, it is plausible that the IRA was successful in influencing U.S. public opinion in

2016.

Introduction

While social media originally allowed a decentralized sharing of information by individuals (Tufekci & Wilson, 2012; Tufekci, 2017), it has more recently provided state actors with new tools for propaganda (Gunitsky, 2015; Rød & Weidmann, 2015; Spaiser, Chadefaux, Donnay, Russmann, & Helbing, 2017; Sanovich, 2017). The extent to which bad information propagates on social media (Vosoughi, Roy, & Aral, 2018) has led to speculation that disinformation on social media can affect U.S. political opinion and even election outcomes (Benkler, Faris, & Roberts, 2018; Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019; Allcott & Gentzkow, 2017).

The Russian Internet Research Agency (IRA) has sought to influence public opinion in many countries (Spaiser et al., 2017; Grigas, 2016; Narayanan, Howard, Kollanyi, & Elswah, 2017; Neudert, Kollanyi, & Howard, 2017; Bodine-Baron, Helmus, Radin, & Treyger, 2018) using state directed disinformation campaigns (Lysenko & Brooks, 2018). In fact, the phrase "Kremlin Troll" has become a term of abuse in Lithuanian comment threads (Zelenkauskaite & Niezgoda, 2017). As stated on page 4 of the Mueller Report (Mueller, 2019) released in April 2019, the IRA carried out "a social media campaign designed to provoke and amplify political and social discord in the United States." There is debate, however, as to how substantially IRA disinformation affected public opinion leading up to the 2016 U.S. presidential election (Jamieson, 2018; Badawy, Ferrara, & Lerman, 2018; Howard, Kelly, & Camille François, 2019; Guess, Nagler, & Tucker, 2019; Allcott, Gentzkow, & Yu, 2019; Garrett, 2019). Here we test a condition of this hypothesis, in whether IRA Twitter activity predicted future changes in 2016 election opinion polling—'predicted' meaning that information in one time series contains information about the future activity in other time series. Causation is not proven by this analysis, but certain directions of causality can be ruled out when one time series does not predict the other.

We take the view that IRA Twitter activity was representative of a larger, multi-media disinformation campaign (Paul & Matthews, 2016; Schoen & Lamb, 2012; Benkler et al., 2018). Here we apply Vector Auto Regression (VAR) to compare weekly time series of re-tweet and 'like' activity of IRA tweets versus weekly data from U.S. election opinion polls in 2016.

In October 2018, Twitter released the content of "Twitter accounts potentially connected to a propaganda effort by a Russian government-linked organization known as the Internet Research Agency" to both the United States Congress and the public (Twitter, 2017).

The Twitter data contains over 9 million tweets representing the activity of 3,613 IRA linked accounts (Twitter, 2017). Of these, 770,005 English language tweets occurred during the 2016 presidential campaign. The number of tweets per week increased during the campaign (see figure 2A). To correct for this, we used 'number of retweets per tweet' as our first measure of success. We confirm our findings using a second measure: 'number of likes per tweet'. We also see that 91% of first retweeters of IRA tweets were non-IRA bots, which suggests that propaganda spread into networks of real U.S. citizens.

The opinion polling data come from Fivethrtyeight.com (FiveThirtyEight, 2017). Fivethirtyeight compiled a database of 3315 national polls from 54 pollsters asking whether the participant intended to vote for Donald Trump or Hillary Clinton; many also included Gary Johnson as a third option. The Fivethirtyeight data exist in two forms, raw weekly time series of Trump and Clinton's polling percentage, as well as an adjusted poll that corrects for the presence/absence of a third candidate; likely voter status; smoothing and political bias of the pollster. Time series were built by averaging all national polls in a given week across all pollsters (see Supplement for a list of pollsters).

Our first VAR tests if the weekly time series of retweets for IRA Twitter accounts, R_t, predicted the next week's changes in election polls for Trump T_t and/or Clinton C_t:

$$T_t \sim T_{t-1} + R_{t-1} \tag{1}$$

$$C_t \sim C_{t-1} + R_{t-1} \tag{2}$$

VAR analysis is then repeated, only using 'likes' rather than re-tweets as the measure of R_t. Conversely, we also tested whether polling activity predicted IRA Twitter success, such that T_t or C_t predicted future R_t:

$$R_t \sim R_{t-1} + T_{t-1} \tag{3}$$

$$R_t \sim R_{t-1} + C_{t-1} \tag{4}$$

Akaike Information Criterion indicates that a time-lag of one week is optimum for these VAR tests. The statistical significance of each VAR result is tested by "Granger Causality" (Granger, 1969), a statistical test of prediction rather than true causality.

We run a number of robustness checks (Supplementary Materials) that measure IRA Twitter activity and polling in different ways. We also controlled for the number of re-tweets from Donald Trump's personal Twitter account using data from the 'Trump Twitter Archive' (Brown, 2019) because it is a possible cause for both opinion polls and IRA Twitter activity.

Results

Timeline indicates disinformation strategy

IRA activity on Twitter was an order of magnitude larger than that on other social media platforms: around 1,500 posts per week on Facebook and Instagram, compared to around 15,000 on Twitter (Howard et al., 2019). Figure 2A shows that during the presidential election itself, this increased to above 25,000 posts per week. However, volume of Tweets did not necessarily translate into more retweets or likes per tweet (figure 2C).

The Mueller investigation concluded that Russia attempted to influence U.S public opinion during the 2016 presentational election (Mueller, 2019). Such campaigns have been carried out in other countries too (Grigas, 2016; Spaiser et al., 2017; Narayanan et al., 2017; Neudert et al., 2017; Bodine-Baron et al., 2018) and exhibit a particular modus operandi that is high-volume and multichannel, as well as rapid, continuous and repetitive (Paul & Matthews, 2016; Sanovich, 2017). This MO is evident in the timeline of Twitter bot activity (figure 1a). Firstly, the IRA Twitter activity appears to change abruptly at important political moments. After the San Bernardino shooting (12/02/2015), for example, five new IRA Twitter accounts, including "TEN_GOP" and others, were introduced (figure 1b). In the tweets from these accounts, 'police', 'shooting' and 'muslim' were among the most frequently used words (figure 1b). The IRA Twitter accounts showcased in figure 1 are the most prominent measured by total number of tweets, retweets and followers. Incidentally, the number of re-tweets from these accounts was correlated ($r = 0.8$) and co-evolving (Supplementary Materials) with Trump's personal Twitter activity, suggesting they had a similar audience.

Many of the most prominent IRA Twitter accounts were imitating U.S news sources and tweeted using convincing English (table 1). Among the most successful 25 IRA individual tweets (Supplementary Materials), two prominent themes emerge in-line with the Russian modus operandi (Paul & Matthews, 2016): discrediting an establishment figure in Hilary Clinton and emphasizing pre-existing societal divisions by focusing on black racial identity.

IRA Twitter success predicted election opinion polls

As the popularity of presidential candidates ebbed and flowed during the 2016 campaign (figure 2B), changes in opinion poll numbers for Trump were consistently preceded by corresponding

# Retweets	Containing the word “black”
41,089	When Its slowly becoming illegal for black people to work BlackTwitter
25,169	He didnt want a black nurse to help his dying child And now his child is gone Pathetic and ridiculous
15,376	Black TMZ staffer schools his white coworkers over The Weeknds hair Today in black history
14,664	The Angry Dark Skin Friend Ay this is very important ht
13,316	Wow Hadnt thought of it that way but thats exactly what is happening So true BlackLivesMatter
	Containing the word “Hilary”, “Clinton” or “Trump”
15,548	BREAKING VoterFraud by counting tens of thousands of ineligible mail in Hillary votes being reported in Broward County Florida Please RT
12,716	OMG this new AntiHillary ad is brilliant Its fantastic Spread it far amp wide
10,824	RT the hell out of it Dem party operatives Weve been bussing people in for 50 yrs and were not going to stop now EvangelicalTrump
10,275	Some guy right in Hillarys face HILLARY FOR PRISON Hillary LETS MAKE IT HAPPEN I almost feel sorry for her
9,323	BREAKING Hillary shuts down press conference when asked about DNC Operatives corruption amp VoterFraud debatenight TrumpBookReport

Table 1: **Top:** Top 5 most retweeted tweets containing the word ‘black’. **Bottom:** Top 5 most retweeted tweets containing ‘Hilary’, ‘Clinton’ or ‘Trump’

changes in IRA re-tweet volume, at an optimum interval of one week before (figure 2C&D). Compared to its time-average of about 38%, support for Trump increased to around 44% when IRA tweets were at their most successful (figure 2D).

Vector Auto-Regression (VAR) and Granger Causality tests provide statistical support that IRA twitter success (measured as both retweets/likes per tweet) predicted future increases for Trump in the polls, but did not predict Clinton's polls (figure 3). Conversely, neither set of opinion poll numbers correlated with future re-tweets or 'likes' of the IRA Tweets.

This result proved robust when running the analysis in a number of different ways: measuring twitter success as total number of re-tweets (not the average), shortened time resolution (two days), using polling end date (not start date), using Twitter likes and using adjusted polls. In none of these tests, however, does the raw number of original IRA tweets predict the polls. Instead, it is the re-tweets, not total volume of original IRA tweets, that predicts the opinion polls (see Supplementary Materials for robustness checks and "Granger Causality" tests). We also discovered that IRA retweets still predicted Trump's polls with the same magnitude when we controlled for the possible confounding effect of average weekly re-tweets from Trump's personal Twitter account P_t (see Supplementary Materials).

Overall, the effect is quantified such that a gain of 25,000 re-tweets per week over all IRA tweets (or about 10 extra re-tweets per tweet per week), predicted approximately 1% increase in Donald Trump's poll numbers.

Discussion and Conclusions

Here we have (a) examined the timing of the IRA Twitter activity, which suggests a strategic release in parallel with significant political events before the 2016 election and (b) used Vector Autoregression (VAR) to test if the success of IRA activity on Twitter predicted changes in the 2016 election opinion polls. On a weekly time scale, we find that multiple time series of

IRA tweet success robustly predicted increasing opinion polls for the Republican candidate, Donald Trump, but not the Democratic candidate, Hillary Clinton. The opinion polls do not predict future success of the IRA tweets. The findings proved robust to many different checks, including a control for the average number of retweets for Donald Trump's personal Twitter account.

The measured effect, a 1% poll increase for Donald Trump for every 25,000 weekly re-tweets of IRA messages, raises two questions about the effect: one regarding the magnitude and one regarding the asymmetry favoring Trump.

Here we have tested prediction, not causality. It seems unlikely that 25,000 re-tweets could influence 1% of the electorate in isolation (Guess et al., 2019; Allcott et al., 2019), although this might be more plausible given that only about 4% of viewed tweets result in re-retweets (Lee, Mahmud, Chen, Zhou, & Nichols, 2014), such that 25,000 re-tweets could imply about 500,000 exposures to those messages per week. It is more likely that Twitter is just a subset of a larger disinformation campaign carried out on multiple social media platforms (Issac & Wakabayashi, 2017; Howard et al., 2019), as well as spread through social contagion (Centola, 2010) and to other parts of the interconnected 'media ecosystem' including print, radio and television (Benkler et al., 2018). In this way way IRA disinformation can frame the debate, meaning many more people than those directly exposed can be affected (Jamieson, 2018).

Any correlation established by an observational study could be spurious. Though our main finding has proved robust and our time series analysis excludes reverse causation, there could still be a third variable driving the relationship between IRA Twitter success and U.S. election opinion polls. We controlled for one of these — the success of Donald Trumps personal Twitter account — but there are others that are more difficult to measure; including exposure to the U.S domestic media, such as Fox news, Breitbart, MSNBC etc.

The asymmetrical effect favoring Republican candidate Donald Trump could be because

Republican supporters were targeted by the IRA (Miller, 2019), making them 36 times more likely to retweet IRA content than Democratic supporters (Badawy et al., 2018). Moreover, Republican supporting regions of the U.S media ecosystem were more susceptible to disinformation than Democrat supporting regions (Benkler et al., 2018), meaning increased sentiments of anger and fear around the time of the election (Miller, 2019) may have helped mobilize Republican voters behind Donald Trump.

We use macro-level data to establish a link between exposure to IRA disinformation and changes in U.S. public opinion. However, using aggregated data means we cannot know the extent to which the participants in election polls were exposed to IRA disinformation. This may not matter once social contagion (Centola, 2010) and media ecosystem effects (Benkler et al., 2018) are taken into consideration. Nonetheless, establishing individual-level causal mechanisms should be a priority (Gerber & Zavisca, 2016; Spaiser et al., 2017).

References

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236.

Allcott, H., Gentzkow, M., & Yu, C. (2019, jan). Trends in the Diffusion of Misinformation on Social Media. *National Bureau of Economic Research*.

Badawy, A., Ferrara, E., & Lerman, K. (2018). Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.*.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bodine-Baron, E., Helmus, T., Radin, A., & Treyger, E. (2018). *Countering Russian Social Media Influence*. Santa Monica: RAND Corporation.

Brown, B. (2019). *Trump Twitter Archive.* Retrieved 2019-03-14, from `http://www.trumptwitterarchive.com/about`

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, *329*(5996), 1194–7.

FiveThirtyEight. (2017). *2016 Presidential Elections: National Polls.* Retrieved 2019-02-24, from https://projects.fivethirtyeight.com/2016-election-forecast/national-po

Garrett, R. K. (2019). Social media's contribution to political misperceptions in U.S. Presidential elections. *Plos One*, *14*(3).

Gerber, T. P., & Zavisca, J. (2016). Does Russian Propaganda Work? *The Washington Quarterly*, *39*(2), 79–98.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424.

Grigas, A. (2016). *Beyond Crimea: The New Russian Empire*. New Haven: Yale University Press.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1).

Gunitsky, S. (2015, mar). Corrupting the Cyber-Commons: Social Media as a Tool of Autocratic Stability. *Perspectives on Politics*, *13*(01), 42–54.

Howard, P. N., Kelly, J., & Camille François, G. (2019). The IRA, Social Media and Political Polarization in the United States, 2012-2018. *Project on Computational Propaganda*.

Issac, M., & Wakabayashi, D. (2017, oct). Russian Influence Reached 126 Million Through Facebook Alone. *The New York Times*. Retrieved from https://www.nytimes.com

Jamieson, K. H. (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President*. Oxford: Oxford University Press.

Lee, K., Mahmud, J., Chen, J., Zhou, M., & Nichols, J. (2014). *Who Will Retweet This? Automatically Identifying and Engaging Strangers on Twitter to Spread Information.*

Lysenko, V., & Brooks, C. (2018, apr). Russian information troops, disinformation, and democracy. *First Monday*, *22*(5).

Miller, D. T. (2019, apr). Topics and emotions in Russian Twitter propaganda. *First Monday*, *24*(5).

Mueller, R. S. (2019). *Report On The Investigation Into Russian Interference In The 2016 Presidential Election* (Tech. Rep.). Washington: U.S. Department of Justice.

Narayanan, V., Howard, P. N., Kollanyi, B., & Elswah, M. (2017). Russian Involvement and Junk News during Brexit. *Computational Propaganda Project*. Retrieved from http://blogs.oii.ox.ac.uk

Neudert, L.-M., Kollanyi, B., & Howard, P. N. (2017). Junk News and Bots during the German Parliamentary Election: What are German Voters Sharing over Twitter? *Computational Propaganda Project*. Retrieved from https://comprop.oii.ox.ac.uk/

Paul, C., & Matthews, M. (2016). *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*. Santa Monica: RAND Corporation.

Rød, E. G., & Weidmann, N. B. (2015). Empowering activists or autocrats? The Internet in authoritarian regimes. *Journal of Peace Research*, *52*(3), 338–351.

Sanovich, S. (2017). *Computational Propaganda in Russia: The Origins of Digital Misinformation* (Tech. Rep.). Retrieved from `https://comprop.oii.ox.ac.uk/`

Schoen, F., & Lamb, C. J. (2012). *Deception, Disinformation, and Strategic Communications: How One Interagency Group Made a Major Difference*. National Defense University Press,. Retrieved from `https://searchworks.stanford.edu/`

Spaiser, V., Chadefaux, T., Donnay, K., Russmann, F., & Helbing, D. (2017). Communication power struggles on social media: A case study of the 201112 Russian protests. *Journal of Information Technology & Politics*, *14*(2), 132–153.

Tufekci, Z. (2017). *Twitter and Tear Gas: the Power and Fragility of Networked Protest*. Yale University Press.

Tufekci, Z., & Wilson, C. (2012). Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. *Journal of Communication*, *62*(2), 363–379. doi: =

Twitter. (2017). *IRA twitter data.* Retrieved 2019-02-24, from `https://blog.twitter.com/`

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151.

Zelenkauskaite, A., & Niezgoda, B. (2017, may). Stop Kremlin trolls: Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting. *First Monday*, *22*(5).

Acknowledgments

Special thank you to Suzie Allard for her helpful comments on early drafts

DJR is funded by the College of Arts and Sciences and the Office of Research and Engagement at the University of Tennessee

NMR is funded by a Minerva grant (Department of Defense) and the Office of Research and Engagement at the University of Tennessee

The authors declare that they have no competing financial interests.

Figures

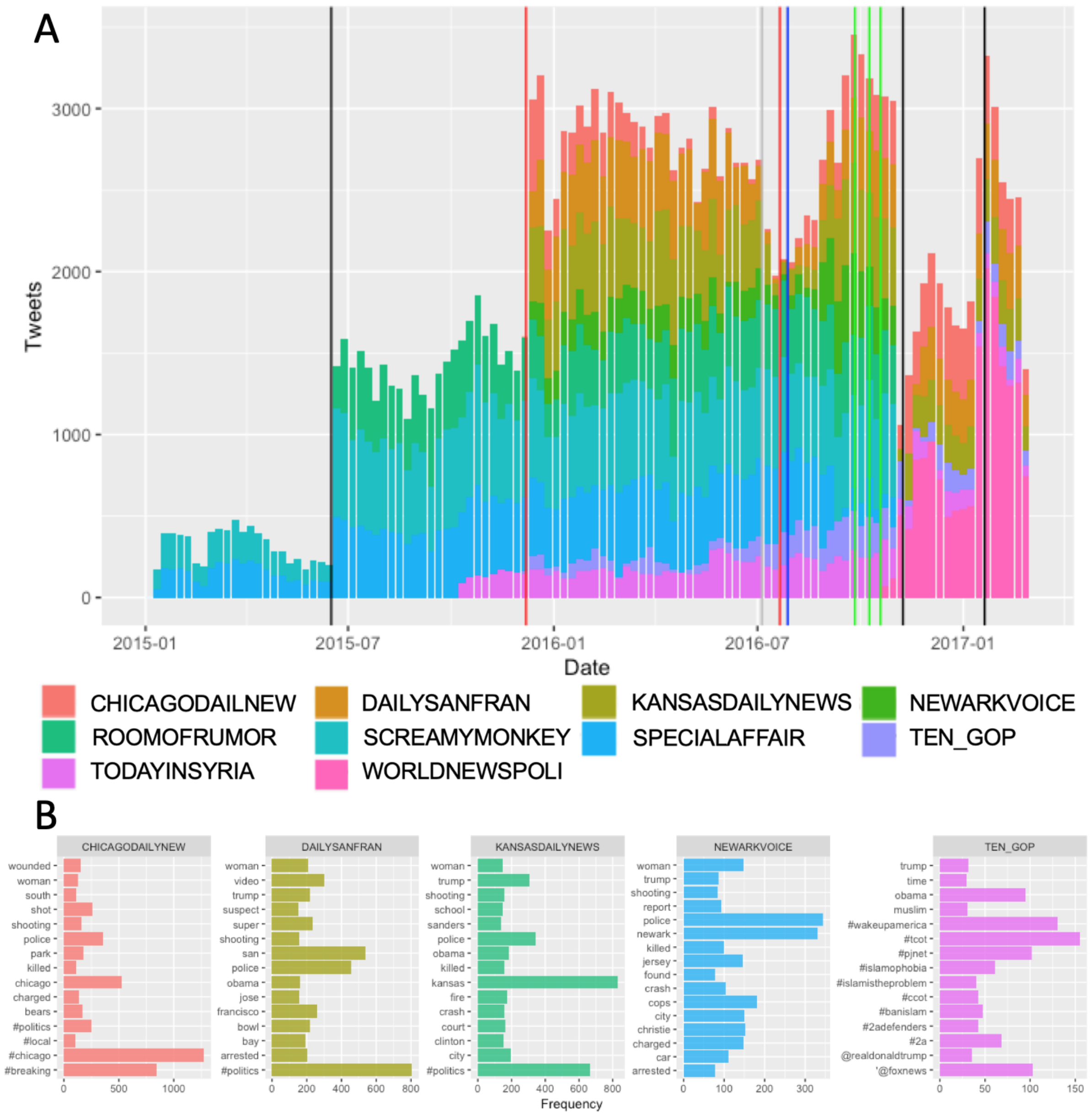

Figure 1: (A) number of weekly tweets by the 10 most prominent IRA Twitter accounts, with vertical lines indicating important offline events during the election cycle (from left to right: Trump's candidacy (black, 06/16/2015), San Bernardino shooting (red, 12/07/2015), Trumps declared nominee (red,06/19/2016), Hillary declared nominee (blue, 07/28/2016), primary debates (green, 09/26, 10/09, 10/19/2016), election day (black, 11/08/2016), and Trumps inauguration (black, 01/20/2017)). (B) most commonly tweeted words from the five prominent IRA Twitter accounts created after the San Bernardino shooting.

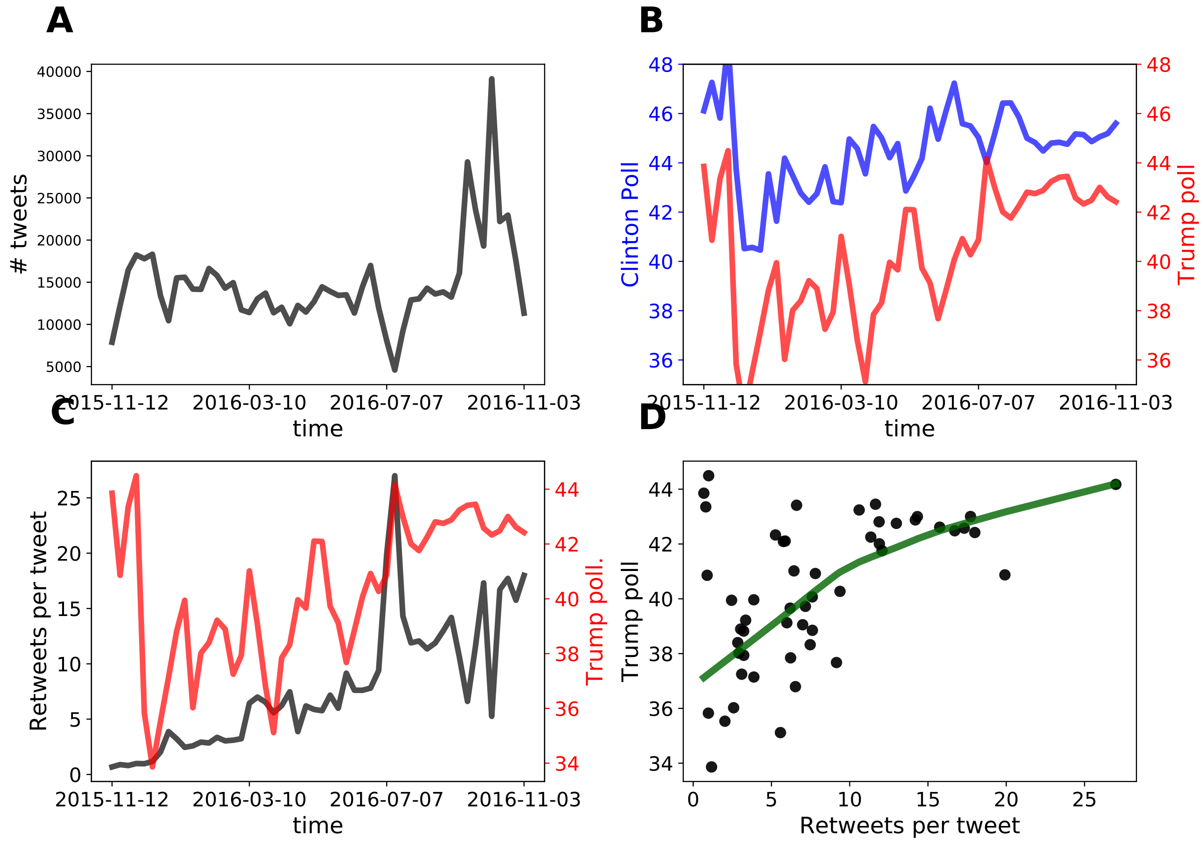

Figure 2: (A) number of IRA tweets per week during election campaign. (B) Trump and Clinton's polling (averaged over 3315 polls). (C) time series for IRA 're-tweets per tweet' and Trump's polls. (D) LOESS fit of contemporaneous IRA 're-tweets per tweet' and Trump's polls

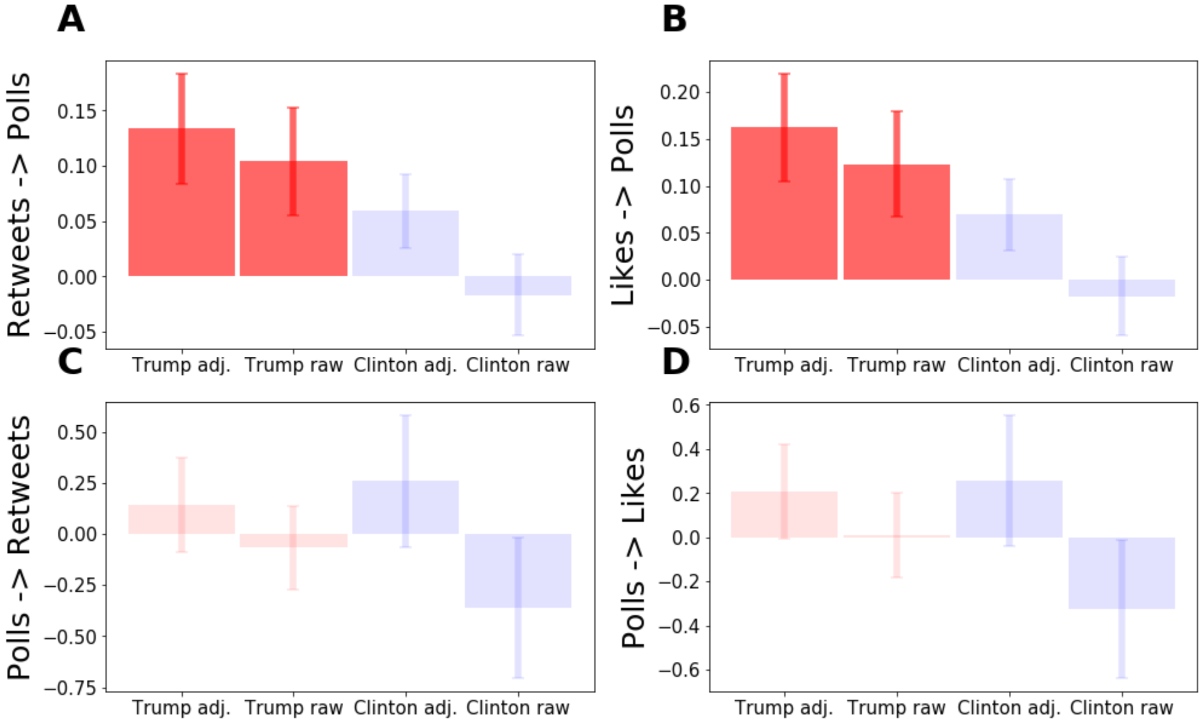

Figure 3: Vector Auto-Regression (VAR) showing IRA twitter success predicts future increases in Donald Trump's polling. Granger causality of Trump and Clinton's polls by IRA (A) retweets per tweet and (B) likes per tweet; (C) and (D) test for the reverse Granger causation. Height of bars are effect sizes, error bars are standard errors and bars are opaque if statistical significance was attained ($p < 0.05$).