Volume 6 | Spring 2019

DEFENCE STRATEGIC COMMUNICATIONS

The official journal of the
NATO Strategic Communications Centre of Excellence

ISSN: 2500-9486
DOI: 10.30966/2018.RIGA.6

MEASURING THE EFFECT OF RUSSIAN INTERNET RESEARCH AGENCY INFORMATION OPERATIONS IN ONLINE CONVERSATIONS

John D. Gallacher and
Marc W. Heerdink

Abstract

The Internet has given new opportunities to those who wish to interfere and disrupt society through the systematic manipulation of social media. One goal of these cyber-enabled information operations is to increase polarisation in Western societies by stoking both sides of controversial debates. Whether these operations are successful remains unclear. This paper describes how novel applications of computational techniques can be used to test the impact of historical activity from the Russian Internet Research Agency (IRA) on two social media platforms: Twitter and Reddit. We show that activity originating from the Russian IRA had a measurable effect on the subsequent conversations of genuine users. On Twitter, increases in Russian IRA activity predicted subsequent increases in the degree of polarisation of the conversation surrounding the Black Lives Matter movement. On Reddit, comment threads started by Russian IRA accounts contained more toxic language and identity-based attacks. We use causal analysis modelling to further show that Russian IRA activity in existing threads caused measurable changes in the conversational quality of the following 25-100 posts. By developing methods to measure the impact of information operations in online conversations and demonstrating a measurable effect on genuine conversations, our study provides an important step in developing effective countermeasures.

Keywords—*information operations, social media, social psychology, group polarisation, disinformation, strategic communications*

About the Authors

John D. Gallacher is a Cyber Security DPhil student at the University of Oxford working with the department of Experimental Psychology and the Oxford Internet Institute.

Dr Marc W. Heerdink is an Assistant Professor of Social Psychology at the University of Amsterdam, where he also obtained his PhD, and a former visiting postdoc at the University of Oxford.

***

Introduction

The rapid development of the Internet has enabled people everywhere to connect, communicate, and distribute information globally at an unprecedented scale. However, some use this opportunity for connection to divide rather than to bring people together. In recent years, a great deal of attention has been focused on groups that conduct deliberate social media activities to divide and polarise societies. These activities include the use of artificial social media accounts, paid advertisements, and automated scripts designed to spread disinformation.[1] These activities are constituents of wider information operations campaigns that seek to gain a competitive international advantage over traditional adversaries.[2] While the approach itself is not new—similar methods targeting the psychology of civilian populations can be traced back to the Roman, Persian, and Chinese empires[3] —these methods have transformed in the digital age and now increasingly rely on social media platforms that provide global reach and can target individuals directly for a fraction of the cost of traditional methods.[4] This phenomenon is characterised by sustained and pervasive efforts[5], which

[1] Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan, 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', The William and Flora Hewlett Foundation, 2018.
[2] Christopher Paul and Miriam Matthews, "'The Russian 'Firehose of Falsehood' Propaganda Model," 2016
[3] Jen Weedon, William Nuland, and Alex Stamos, *Information Operations and Facebook*, Facebook, 2017.
[4] Edward Lucas and Peter Pomerantsev, 'Winning the Information War: Techniques and Counter-Strategies to Russian Propaganda in Central and Eastern Europe', Center for European Policy Analysis & The Legatum Institute, 2016.
[5] John D Gallacher and Rolf E Fredheim, 'Division Abroad, Cohesion at Home: How the Russian Troll Factory Works to Divide Societies Overseas but Spread pro-Regime Messages at Home', in Responding to Cognitive Security Challenges (Riga, Latvia: NATO StratCom CoE, 2019), 60–79.

peak around election cycles, although elections are not the sole focus. This persistent engagement, short of traditional thresholds for conflict, makes it difficult to construct robust responses.[6]

In 2014, the World Economic Forum identified the rapid spread of misinformation online as one of the top 10 threats to society.[7] Since this warning, the deliberate spread of misleading information has been linked to political earthquakes such as the 2016 US Election[8], the 2016 UK Brexit referendum[9], and the rise of populist parties across Europe[10], as well as to political violence in Brazil[11], Myanmar[12], and India[13]. All these events are connected by one consistent trend—an increase in social polarisation, defined as the process of increased segregation into distinct social groups, separated along racial, economic, political, religious or other lines.[14] Hostile information operations on social media show no evidence of slowing down[15], while social media platforms stand accused of failing to act decisively in combatting this threat[16]. Understanding the consequences of these activities is essential to developing effective defences. In-depth knowledge about the consequences of these hostile narratives should inform policy decisions aimed at countering them, yet very little is known about the effect these activities have on the online conversations of genuine citizens, and whether or not they achieve their goals.

In this study we developed methods to address this question and to measure the effect of artificial social media manipulation on subsequent human conversations, using publicly attributed information operations from the Russian state as a case study. Recently, evidence shows that the Russian government has been engaged in a substantial effort to sway public opinion on a number of

[6] Lucas Kello, *The Virtual Weapon and International Order* (Yale University Press, 2017).
[7] Lee Howell et al., 'Outlook on the Global Agenda 2014' (Geneva: World Econoomic Forum, 2014).
[8] Intelligence Community Assesment, 'Assessing Russian Activities and Intentions in Recent US Elections', Office of the Director of National Intelligence, 2017.
[9] 'Disinformation and "Fake News": Interim Report', UK Department for Digital, Culture, Media and Sport Committee, 2018.
[10] Emilio Ferrara, 'Disinformation and social bot operations in the run up to the 2017 French presidential election', First Monday, Vol. 22, № 8, 2017.
[11] Dan Arnaudo, 'Computational Propaganda in Brazil: Social Bots during Elections', University of Oxford Computational Propaganda Research Project 8 (2017): 1–39.
[12] Steve Stecklow, 'Why Facebook Is Losing the War on Hate Speech in Myanmar', Reuters, 15 August 2018.
[13] Neeta Rani, 'Social Media in India: A Human Security Perspective', The Research Journal of Social Sciences Vol. 9, № 10 (2018): 43–52.
[14] Alisdair. Rogers, Noel Castree, and Rob Kitchen, A Dictionary of Human Geography (Oxford University Press, 2013).
[15] Kanisk Karan, Donara Barojan, Melissa Hall, and Graham Brookie, '#TrollTracker: Outward Influence Operation from Iran', Medium, That Atlantic Council's Digital Foresnsics Research Lab, 31 January 2019.
[16] 'Disinformation and "Fake News": Final Report', UK Department for Digital, Culture, Media and Sport Committee, 2019.

key topics, at home and abroad, through a prolonged information campaign.[17] This campaign includes disinformation, artificial social media accounts imitating a grass-roots movement, paid advertisements, and automated scripts designed to hijack filtering algorithms in order to disseminate content to the widest possible audience.[18] These accounts also promoted real-world protests and demonstrations, often encouraging both sides of controversial topics. While the 2016 US presidential election seems to have been one important focus for these activities, the wider intention appears to have been to polarise online conversations and sow social division along social as well as political lines.[19]

The relationship between disinformation and polarisation

People increasingly use social media as their primary source for news and information, with two-thirds of Americans and half of adults in the developing world getting their news from social media platforms.[20] Ideological alignment with specific groups and ideas is often more obvious in online environments than it is offline,[21] either due to structural features, such as profile pictures or group memberships, or because of the content shared by users. For this reason, separation into groups of likeminded people is more likely to occur online than offline. This facilitates group polarisation, a social-identity-based phenomenon where individuals endorse more extreme ideological positions following a discussion with other in-group members.[22] This increased polarisation may encourage group members to take a more extreme position on certain issues, or may result in an increased dislike of members of other groups without a change in their position on that issue.[23]

[17] Weedon et al., *Information Operations and Facebook*; Intelligence Community Assesment, 'Assessing Russian Activities'.

[18] Renee DiResta, Jonathan Albright, and Ben Johnson, 'The Tactics & Tropes of the Internet Research Agency', New Knowledge Disinformation Report Whitepaper, 2018; Philip N Howard, Bharath Ganesh, Dimitra Liotsiu, John Kelly, Camille Françoise 'The IRA, Social Media and Political Polarization in the United States, 2012-2018', University of Oxford Computational Research Project, 2018.

[19] Sebastian Bay et al., Responding to Cognitive Security Challenges, (Riga, Latvia: NATO StratCom CoE, 2019); DiResta et al., 'Tactics & Tropes'.

[20] Elisa Shearer and Jeffrey Gottfried, 'News Use across Social Media Platforms 2017', Pew Research Center, 17 September 2017; Nic Newman with Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy and Rasmus Kleis Nielsen, 'Reuters Institute Digital News Report 2017' (Reuters Institute for the Study of Journalism, 2017).

[21] Tom Postmes, Russell Spears, and Martin Lea, 'Building or Breaching Social Boundries? SIDE Effects of Computer Mediated Communication', Communication Research 25, № 6 (1998): 689–715; Eun Ju Lee, 'Deindividuation Effects on Group Polarization in Computer-Mediated Communication: The Role of Group Identification, Public-Self-Awareness, and Perceived Argument Quality', Journal of Communication Vol. 57, Issue 2 (2007): 385–403.

[22] J. C. Turner, B. Davidson, and M. A. Hogg, 'Polarized Norms and Social Frames of Reference: A Test of the Self-Categorization Theory of Group Polarization', *Basic and Applied Social Psychology* Vol. 11, № 1 (1990): 77–100.

[23] Lilliana Mason, '"I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization', *American Journal OfPolitical Science* Vol. 59, Issue 1 (2014): 128–45.

Messages emphasising inter-party conflict have been shown to reinforce social polarisation and are easy to distribute in online environments. Messages containing strong partisan cues that match an individual's beliefs can encourage them to accept and share inaccurate information,[24] while messages that agree with pre-held stereotypes can facilitate an individual's acceptance of inaccurate information about an out-group.[25] Equally, polarised conversations can lead to increased dissemination of disinformation. People are more likely to trust inaccurate information if it elicits anger and aligns with their existing opinions.[26] Content that is highly controversial or elicits greater moral outrage is more likely to be shared by social media users,[27] while erroneous content can be made more sensational than true content and therefore more likely to inspire fear and disgust, which in turn encourages sharing the content faster and farther.[28]

Online environments may create 'echo chambers'—networks of like-minded people who confirm each other's opinions instead of promoting critical thinking[29]—exacerbating these effects. Disinformation spreads more quickly within these closely connected groups due to a lack of dissenting voices.[30] This may facilitate the creation of a society that is increasingly polarised and misinformed[31] as people are more likely to be affected by inaccurate information if they see it more frequently, especially in cases where fresh exposure influences decision-making.[32]

[24] R. Kelly Garrett, Brian E. Weeks, and Rachel L. Neo, 'Driving a Wedge Between Evidence and Beliefs: How Online Ideological News Exposure Promotes Political Misperceptions', *Journal of Computer-Mediated Communication* Vol. 21, Issue 5 (2016): 331–48.

[25] James N. Druckman, Erik Peterson, and Rune Slothuus, 'How Elite Partisan Polarization Affects Public Opinion Formation', *American Political Science Review* Vol. 107, Issue 01 (2013): 57–79; R. Kelly Garrett, Shira Dvir Gvirsman. Benjamin K. Johnson, Yariv Tsfati, Rachel Neo, and Aysenur Dal, 'Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization', *Human Communication Research* Vol. 40, Issue 3 (2014): 309–32; Brian E. Weeks, 'Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation', *Journal of Communication* vol. 65, Issue 4 (2015): 699–719; Spee Kosloff, Jeff Greenberg, Toni Schmader, Mark Dechesne, and David Weise, 'Smearing the Opposition: Implicit and Explicit Stigmatization of the 2008 U.S. Presidential Candidates and the Current U.S. President', *Journal of Experimental Psychology*: General Vol. 139, № 3 (2010): 383–98.

[26] Weeks, 'Emotions, Partisanship, and Misperceptions'.

[27] William J. Brady, Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel, 'Emotion Shapes the Diffusion of Moralized Content in Social Networks', *Proceedings of the National Academy of Sciences* Vol. 114, № 28 (2017): 7313–18.

[28] Soroush Vosoughi, Deb Roy, and Sinan Aral, 'The Spread of True and False News Online', *Science* Vol. 359, Issue 6380 (2018): 1146–51; M. J. Crockett, 'Moral Outrage in the Digital Age', *Nature Human Behaviour* Vol. 1 (2017):769–71.

[29] M Conover, J Ratkiewicz, and M Francisco, 'Political Polarization on Twitter', *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Conference Paper (2011): 89–96; Sarita Yardi and Danah Boyd, 'Dynamic Debates: An Analysis of Group Polarization over Time on Twitter', *Bulletin of Science, Technology & Society* 30, № 5 (2010): 316–27.

[30] Eli Pariser, *The Filtter Bubble: What the Internet Is Hiding from You* (New York: Penguin Press, 2011).

[31] Cass R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton University Press, 2017).

[32] Adam J. Berinsky, 'Rumors and Health Care Reform: Experiments in Political Misinformation', *British Journal of Political Science* Vol. 47, Issue 2 (2017): 241–62; Gordon Pennycook, Tyrone D Cannon, and David G Rand, 'Prior Exposure Increases Perceived Accuracy of Fake News', *Journal of Experimental Psychology* Vol. 147, № 12 (2018): 1865–80.

Recent evidence suggests that echo-chambers may not be forming as often as first expected[33], and users are, in fact, exposed to more cross-cutting information online than they would select purely based on choice.[34] Even so, this cross-cutting information may not have a positive effect. Users with more extreme ideological positions are more active on social media[35] and exposure to opposing views online can also increase polarisation by highlighting areas of disagreement.[36] Both situations provide opportunities for those who wish to leverage the polarising effects of social media, either through infiltrating echo chambers to spread negative messages about an out-group without opposition, or by engaging with someone while posing as an out-group member in order to antagonise and create a negative impression of the out-group as a whole.

The St. Petersburg Troll Farm and Online Polarisation

From as early as 2012, information operations conducted over social media have been targeting citizens in the West.[37] These operations originate from the St Petersburg 'troll farm' run by the Russian Internet Research Agency (Russian IRA). The agency aims to influence online conversations about regional, national, and international issues that affect Russian foreign and domestic policy interests.[38] Online manipulation can take the form of 'trolling' orchestrated from human-controlled accounts or political communications spread—by automated accounts (bots).[39] Since 2012, these campaigns have grown steadily in number and scale,[40] and have gained much international attention, particularly surrounding the 2016 US presidential election.[41]

[33] Pablo Barberá, John T. Jost, Johnathan Nagler, Joshua A. Tucker, Richard Bonneau, 'Tweeting From Left to Right: Is Online Political Communication More than an Echo Chamber?', *Psychological Science* Vol. 26, № 10 (2015): 1531–42; Jonathan Bright, 'Explaining the Emergence of Echo Chambers on Social Media: The Role of Ideology and Extremism', *SSRN Electronic Journal*, (2016).

[34] Etyan Bakshy, Solomon Messing, and Lada A. Adamic, 'Exposure to Ideologically Diverse News and Opinion on Facebook', *Science* Vol. 348, Issue 6239 (2015): 1130–32.

[35] Pablo Barberá and Gonzalo Rivero, 'Understanding the Political Representativeness of Twitter Users', *Social Science Computer Review* Vol. 33, № 6 (2015), 712–29; Daniel Preoţiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle Ungar, 'Beyond Binary Labels: Political Ideology Prediction of Twitter Users', *Proceedings Ofthe 55th Annual Meeting of the Association for Computational Linguistics*, 2017, 729–40.

[36] Christopher Bail, Lisa Argyle, Taylor Brown, John Bumpus, Haohan Chen, M.B. Hunzaker et al., 'Exposure to Opposing Views Can Increase Political Polarization: Evidence from a Large-Scale Field Experiment on Social Media', *Proceedings of the National Academy of Sciences*, 2018, 1–6.

[37] Howard et al., 'The IRA, Social Media and Political Polarization'.

[38] Theodore P. Gerber and Jane Zavisca, 'Does Russian Propaganda Work?', *The Washington Quarterly* Vol. 39, Issue 2 (2016): 79–98; Sergey Sanovich, 'Computational Propaganda in Russia: The Origins of Digital Misinformation', University of Oxford Computational Propaganda Research Project, 2017.

[39] Samuel C Woolley and Philip N Howard, 'Political Communication, Computational Propaganda, and Autonomous Agents' *International Journal of Communication*, 10 (2016): 4882–90; Rolf Fredheim, 'Robotrolling 2019, Issue 1' (Riga, Latvia, NATO StratCom COE, 2019); Srijan Kumar, Justin Cheng, Jure Leskovec, V. S. Subrahmanian, An Army of Me: Sockpuppets in Online Discussion Communities', *Proceedings of the 26th International Conference on World Wide Web* (2017), 857–66.

[40] Howard et al., 'The IRA, Social Media and Political Polarization'; DiResta et al., 'Tactics & Tropes'.

[41] Intelligence Community Assesment, 'Assessing Russian Activities'.

Over the course of 2018, large, open-source datasets detailing posts from accounts attributed to the Russian IRA were published, making it possible to conduct a detailed analysis of how Russia ran these information campaigns.[42] The data showed that the campaign was not restricted to the 2016 US election but rather sought to divide online groups along racial, ethnic, social, and political lines, and continued long after the election was decided.[43] Both sides of numerous controversial debates were inflamed by Russian IRA activity, especially conversations surrounding provocative race issues such as the Black Lives Matter movement in the United States.

Measuring the effect of these information operations

While the intention behind this activity is clear, measuring its impact is complex. Trolls have been shown to manipulate the opinions of users in online forums[44] and to steer conversations on blogging platforms.[45] While at times these accounts have garnered greater popularity than those of organic users,[46] the impact they have on the wider online ecosystem is hard to measure. Some calculations show that Russian IRA accounts were influential in spreading targeted URLs across Twitter,[47] but that this activity did not carry over to other web communities (Reddit, 4Chan).[48] Twitter's key role in these campaigns is also illustrated by the fact that in the run-up to the 2016 US Election, more hyperlinks to websites hosting disinformation were shared on Twitter than across the top sixteen mainstream media outlets combined.[49] What is not clear from this evidence however, is what effect the Russian IRA accounts have had on more subtle areas such as promoting ideas in line with Russian interests, engaging other users to sway opinion, and fuelling both sides of controversial online discussions.

[42] Darren L Linvill and Patrick L Warren, 'Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building', (*in press*), 2018; Matthew Hindman and Vlad Barash, 'Disinformation, "Fake News" and Influence Campaigns on Twitter', 2018.

[43] Gallacher and Fredheim, 'Division Abroad, Cohesion at Home'; Linvill and Warren, 'Troll Factories'.

[44] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov, 'Finding Opinion Manipulation Trolls in News Community Forums', *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, № July (2015): 310–14.

[45] Anton Sobolev, 'Fantastic Beasts and Whether They Matter: How pro-Government "Trolls" Influence Political Conversations in Russia', (*In Prep*).

[46] Howard et al., 'The IRA, Social Media and Political Polarization'.

[47] Savvas Zannettou et al., 'Who Let the Trolls out? Towards Understanding State-Sponsored Trolls', (2019); Savvas Zannettou et al., 'Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter', (2019).

[48] Savvas Zannettou et al., 'Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web', (2018).

[49] Pablo Barberá, 'Explaining the Spread of Misinformation on Social Media: Evidence from the 2016 U.S. Presidential Election', *Comparative Politics Newsletter* Vol. 28, Issue 2 (2018): 7–11.

In this paper we use a two-part strategy to measure the effect of information operations on online conversations. In Part 1 we focus on a case study of the Black Lives Matter (BLM) movement which was targeted by Russian IRA accounts. This social movement has spread both online and offline to protest the systematic violence perpetrated against African-Americans, particularly by police officers.[50] Opposition movements to BLM (#BlackLivesMatter) have criticised it for failing to appreciate the value of all races (#AllLivesMatter) or for failing to respect the value of police officers and the risk they take in course of their work (#BlueLivesMatter).[51] These hashtags can shape how information flows through the wider network and therefore play a significant role in the spreading of ideas.[52] Russian IRA accounts imitated authentic users on both sides of this debate to disseminate provocative messages to various target audiences and to foster antagonism between opposing groups.[53] This is likely to have contributed to the polarisation of the #BlackLivesMatter conversation online; Russian IRA accounts were in the top percentile of retweeted accounts in both supporting and opposing sides of the Twitter conversation.[54] We investigated the global effect of the Russian IRA tweets on the entire #BlackLivesMatter conversation by testing whether the daily degree of polarisation of the Twitter conversation correlated positively with earlier Russian IRA activity surrounding the #BlackLivesMatter hashtag.

In Part 2 we look at the impact of Russian IRA activity on Reddit using natural language programming, text analysis measures, and causal impact modelling to analyse the effect of >16,000 Reddit posts attributed to the Russian IRA. Following revelations about the scope of Russian IRA manipulation of social media platforms in 2016, Reddit was the only social media company to keep this activity publicly visible on the platform rather than removing it, so it is the only platform where the immediate response to Russian IRA content can be analysed directly. We measure the response to known artificial activity and predict that

[50] Deen Freelon, Charlton D. McIlwain, and Meredith Clark, 'Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice', *SSRN Electronic Journal*, (2016).

[51] Leo G. Stewart, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird, 'Drawing the Lines of Contention: Networked Frame Contests within #BlackLivesMatter Discourse', *Proceedings of the ACM on Human-Computer Interaction* Vol. 1, Issue CSCW, Article № 96 (2017): 1–23.

[52] Leo G. Stewart, Ahmer Arif, and Kate Starbird, 'Examining Trolls and Polarization with a Retweet Network', *Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*., 2018; Ryan J. Gallagher et al., 'Divergent Discourse between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter', *PLoS ONE* 13, № 4 (2018): 1–23.

[53] Ahmer Arif, Leo G Stewart, and Kate Starbird, 'Acting the Part: Examining Information Operations within #BlackLivesMatter Discourse', *Proceedings of the ACM on Human-Computer Interaction* Vol. 2, Issue CSCW, Article № 20 (2018): 1–26.

[54] Stewart et al., 'Examining Trolls and Polarization'.

Russian IRA activity causes a measurable decrease in the quality on discussion threads.

In this study we do not make any attributions to which accounts were operated from the Russian IRA. Instead the accounts were identified and attributed by the social media platforms themselves using information that is not available to the public.

Methods

How does the degree of daily polarisation of the #BlackLivesMatter conversation on Twitter correlate with Russian IRA activity?

Data collection and sampling

Twitter is a popular social media platform built on a microblogging format. Users can share short messages, or 'tweets', with their followers who can in turn 'retweet' these messages to their own followers. Tweets can sometimes contain hashtags indicating that it is part of a broader conversation. In late 2018 Twitter averaged 321 million active monthly users.[55]

We obtained Twitter data relating to the Black Lives Matter conversation from an archive complied by the digital chronicling organisation 'Documenting the Now' (DocNow).[56] The dataset contains 17,292,130 tweet IDs for tweets collected from the Twitter streaming API for #BLM and #BlackLivesMatter between 29 January 2016 and 18 March 2017.[57] Twitter's terms of service do not allow public redistribution of tweets; however, they do allow datasets of tweet IDs to be shared. We then recovered the full tweet from each tweet ID by performing a search through the Twitter search API (also known as 'hydration') using DocNow's Hydrator software.[58]

Only tweets which were still publicly available at the time of the search could be recovered; we could not recover tweets that had been deleted by Twitter or by the users themselves.

[55] Statista, 'Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions)', 2019, accessed February 25, 2019.
[56] Documenting the Now.
[57] Ed Summers, Black Lives Matter Tweets 2016, *The Internet Archive*, 17 October 2017.
[58] DocNow Hydrator on GitHub.

We hydrated the dataset of tweet IDs on 24 November 2018, which led to a collection of 9,531,526 tweets, or 55% of available tweet IDs (45% of the original tweets had been deleted since publication). While our dataset, therefore, does not represent the full conversation, it is the best approximation available given the limits that Twitter places on data sharing. Importantly, this dataset does not contain the tweets from Russian IRA, as this information was removed from the platform at the point of attribution by Twitter, prior to collection. Therefore, our measure of polarisation reflects the polarisation of the conversation of genuine (i.e. non-Russian IRA) accounts without potential artificial inflation from Russian IRA tweets.

Data on the activity of known Russian IRA accounts were collected by Linvill and Warren[59], and made publicly available by the team at fivethirtyeight.com.[60] This dataset contains 2,973,371 tweets from 2,848 Twitter accounts spanning the period from 2015–2018.

Measuring polarisation

We measured the degree of daily polarisation on Twitter using a novel technique known as correspondence analysis, implemented in the FactoMineR package for R.[61] Correspondence analysis is a statistical method that makes it possible to map contingency tables to expose underlying relationships between objects in the data.[62] All analyses were performed in R (version 3.4.4, R Core Development Team 2017).

For each day of the dataset, we used a retweet matrix as the contingency table to show the relationship between active users within the dataset (rows) and popular tweets (columns) (see Table 1). A retweet matrix is a good starting point for discovering the structure of Twitter conversations as retweets have been shown to closely represent the expression of agreement with a particular message and, under certain conditions, support of the messenger.[63] Given this, we assumed that if a user retweets messages expressing support or opposition for a given position, this reflects the user's own beliefs.

[59] Linvill and Warren, 'Troll Factories'.
[60] Oliver Roeder, 'Why We're Sharing 3 Million Russian Troll Tweets', *FiveThirtyEight*, 31 July 2018.
[61] Francois Husson et al., 'Package "FactoMineR"', *CRAN*, 2018.
[62] H. O. Hirschfeld and J. Wishart, 'A Connection between Correlation and Contingency', *Mathematical Proceedings of the Cambridge Philosophical Society* Vol. 31, Issue 4 (October 1935): 520.
[63] Panagiotis Metaxas and Twittertrails Research Team, 'Retweets Indicate Agreement, Endorsement, Trust: A Meta-Analysis of Published Twitter Research', *ArXiv Preprint*, 2017.

Popular Tweets ---->

<---- Active Users

	μ^1	μ^2	μ^3	μ^4
ϱ^1	**1**	**2**	0	0
ϱ^2	0	0	**1**	0
ϱ^3	**3**	0	0	**1**
ϱ^4	**1**	**2**	0	0
ϱ^5	0	1	0	**5**

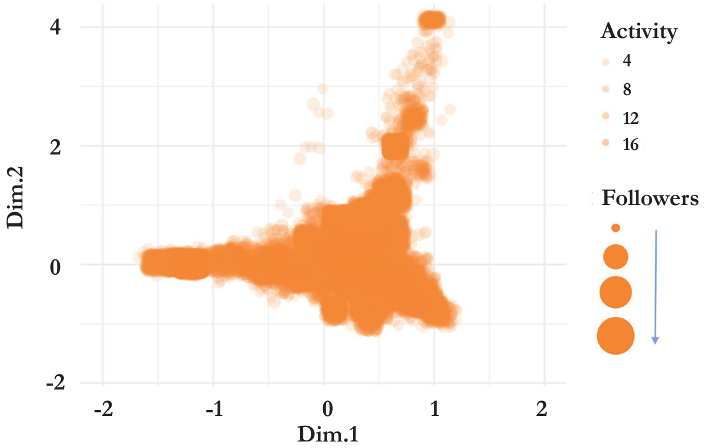

Table 1 and Figure 1 – Simplified retweet matrix for popular tweets and active users for the #BlackLivesMatter Twitter conversation on 07/07/2016 and the correspondence analysis results placing users on dimensions one and two.

Correspondence analysis interprets the retweet matrix across a number of dimensions whereby the largest amount of variability in the data is captured in dimension 1, the next largest amount of variability is captured in dimension 2, the third largest amount of variability is captured in dimension 3. The scores for dimension 1 were used to calculate the position of each tweet on the dimension 1 scale in relation to the other tweets for that day. As explained below, dimension 1 generally distinguishes between tweets that were either for or against the #BlackLivesMatter movement; the greater the score in dimension 1 the stronger the support or criticism. Opposition tweets were often framed as part of a counter-movement, such as #BlueLivesMatter, co-opting BLM-related hashtags (#BlackLivesMater or #BLM) to inject opposing opinions into the conversation.

We focused only on the dimension that demonstrated the greatest variance in the daily activity, dimension 1, because it was the most stable across multiple days and was the most reliable indication of the level of support or opposition for the Black Lives Matter movement indicated by the tweet. We verified the consistency of this dimension by taking a random sample of 50 days from the dataset and selecting the tweets with the highest and lowest scoring days on dimension 1. We manually coded whether the messages presented in these tweets represented opposing sides. This was the case for 85% of the days. Manual inspection of the remaining 15% of days showed that these tended not to have a polarised debate, and so the dimension was absent rather than missed.

To perform a successful a correspondence analysis, the contingency table had to represent a well-connected subgraph of the retweet network to avoid a small subset of users, peripheral to the main conversation, generating large scores on the important dimensions (similar to the k-core within network theory). We therefore used thresholds to filter out less popular tweets (as assessed by the number of retweets) and 'inactive' users (who did not retweet many popular tweets). These thresholds depend on daily conversation size and are shown in Table 2. After ranking all popular tweets along dimension 1, we used the results to estimate the dimension 1 score for each user compared to all other users, based on the average of all the tweets they had retweeted. This last step could be performed for all users, not only those defined as 'active' in the correspondence analysis.

Selecting the correct values for these thresholds is important for achieving stable results. We selected suitable thresholds dynamically for each day according to two rules: (a) thresholds should not produce extreme scores for a subset of users on dimension 1 ($|z| > 10$), and (b) when applying back to scores from the subgraph to all users, thresholds should allow for >25% of daily users in the conversation to be classified as belonging to dimension 1. In rare cases the standard thresholds did not fit; for these days slightly lower/higher thresholds were applied. This was necessary as for some days certain tweets went 'viral', changing the relationship between conversation size and the overall activity of the average user. While setting the thresholds appropriately improved results for each given day, taken overall changing these thresholds did not alter results substantially.

The distribution of users across dimension 1 reflects how their opinions are distributed, and whether users formed distinct 'camps'—something we would expect if the conversation were polarised. We were able to measure the degree of this polarisation using Hartigans' dip test,[64] which measures how bimodal a sample is, with higher scores indicating higher bimodality. We operationalised polarisation as the bimodality of each daily distribution of user scores on dimension 1 (Figure 2).

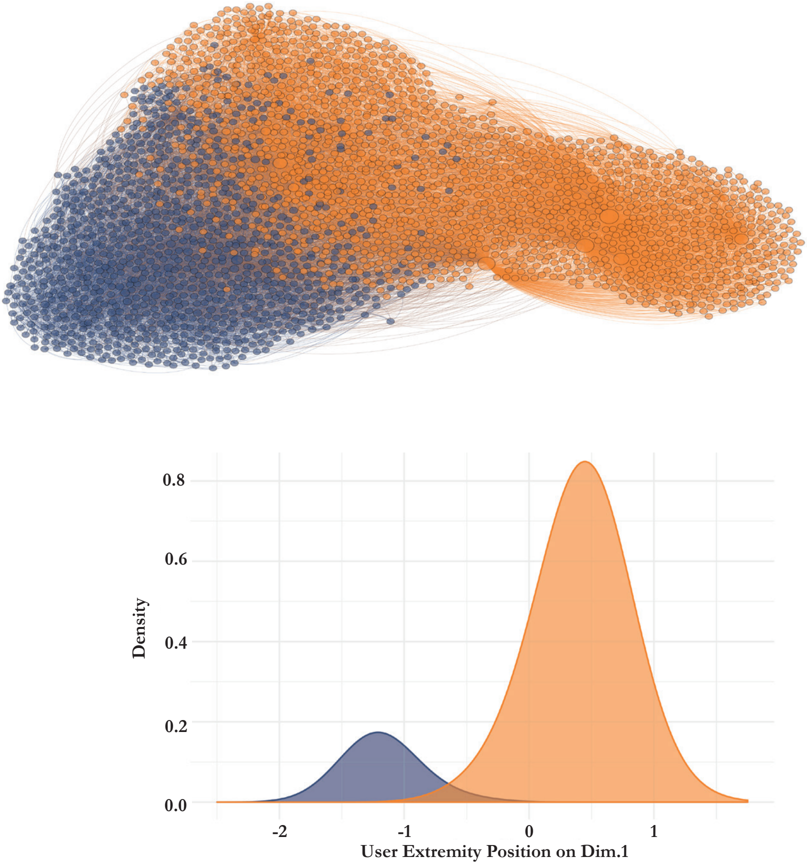

Figure 2(a) and (b) – Visualisation of the polarised retweet network and matching bimodal distribution for dimension 1 for the #BlackLivesMatter conversation on 07/07/2016

[64] Martin Maechler, '<u>Package "Diptest"</u>', *CRAN*, 5 December 2015.

Total conversation size (nb of Tweets)

	>200.000	>100.000	>10.000	>5.000	<5000
Re-Tweet threshold	10	6	4	2	1
Active user threshold	5	3	2	1	1

Table 2 – Thresholds selected for the number of retweets needed for a tweet to be 'popular', and the number of these tweets a user needs to interact with to be considered 'active'.

Relation to Russian Troll Farm Activity

After measuring the degree of polarisation in the daily conversation from genuine accounts, we related it to the artificial activity originating from accounts associated with the Russian IRA using (lagged) Pearson's correlations. Russian IRA activity is measured as the number of posts using a BLM-related hashtag from the public dataset released in summer 2018.

Russian IRA activity is unlikely to have an immediate effect on the degree of polarisation of the conversation, especially as the direct responses to this activity were unavailable. To measure the correlation between Russian IRA activity and the subsequent level of polarisation in the conversation, taking into account cumulative effects of sustained activity over time and delayed effects in the changing dynamics of the conversation, we compared the cumulative Russian IRA activity for a period of 1–7 days prior to each focal day in the dataset with the mean degree of polarisation over the subsequent 1–20 days.

To test if the association between Russian IRA activity and subsequent polarisation was significantly higher than expected by chance, we used a permutation test. For each given level of lag in polarisation (1–20 days) and cumulative period of Russian IRA activity (1–7 days), we simulated a new dataset where Russian IRA activity for each day was paired with a level of polarisation randomly sampled (with replacement) from our real dataset. We then calculated the correlation coefficient between the Russian activity and the lagged polarisation. This was repeated for 10,000 iterations. This circumvented the problem of autocorrelation associated with the lagged time-series as the lagged polarisation was calculated after the randomisation. To avoid biased coefficients arising from right-skewed distributions of activity and polarisation, we normalised the data using box-cox transformations in the R package 'MASS'.[65]

[65] G. E. P. Box and D. R. Cox, 'An Analysis of Transformations', *Journal of the Royal Statistical Society* Vol. 26, № 2 (1964): 211–43; Author Brian et al., 'Package "MASS"', *CRAN*, 2018.

We measured effect sizes for each cumulative period and lag period combination by taking the mean for the total 10,000 simulations. Significance values were calculated as the proportion of simulations where the simulated correlation was higher than the observed correlations.[66]

Measuring the direct effect of Russian IRA activity on Reddit conversations

What is Reddit?

Reddit is a social media platform that focuses on news aggregation and discussion. Content is crowd-sourced, with members submitting text, images, or external hyperlinks, which are then voted up or down by other members. This content is organised into specific 'subreddits', user-created boards covering a wide variety of topics. In February 2018, Reddit had 542 million monthly visitors, ranking as the third most-visited website in the US and the sixth most-visited globally.[67]

Data collection and sampling

Reddit released the identity of Russian IRA accounts in the summer of 2018. This totalled 16,821 Reddit posts from 944 accounts.[68] We extracted our dataset in November 2018 from a publicly available repository of historical Reddit data on pushshift.io.[69] The data are available in the form of a Google Big Query Database, which can be queried by users to download specific sections of the entire database. Here we study the period from January–December 2016, the period during which the Russian accounts were most active. We selected subreddits on which Russian IRA accounts posted at least 50 submissions during 2016. These span a range of topics, allowing us to explore differential effects in different areas of the social media platform. Previous research[70] has demonstrated that some of these subreddits were used by the Russian accounts for political manipulation, while others were used for more mundane purposes such as generating realistic account histories or 'karma' (platform specific credits that give a user more credibility in their comments). We selected the followed

[66] Anthony J. Bishara and James B. Hittner, 'Testing the Significance of a Correlation with Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches', *Psychological Methods* Vol. 17, № 3 (2012): 399–417.
[67] 'Reddit. Com Traffic, Demographics and Competitors—Alexa', accessed 25 February 2019.
[68] 'Reddit's 2017 Transparency Report and Suspect Account Findings', Reddit, accessed 25 February 2019.
[69] https://files.pushshift.io/reddit/comments/
[70] Gallacher and Fredheim, 'Division Abroad, Cohesion at Home'.

DOI 10.30966/2018.RIGA.6.5.

12 subreddits; r/funny, r/uncen, r/Bad_Cop_No_Donut, r/AskReddit, r/PoliticalHumor, r/news, r/worldnews, r/aww, r/gifs, r/politics, r/The_Donald, r/racism. Of these, the subreddit r/uncen had received only submissions from Russian IRA accounts and no comments on the posts and was therefore not included in the analysis. Pushshift collects data at the point that it is posted to Reddit.[71] This means that the dataset is unaffected by subsequent deletion of posts, however it also means that it does not capture edits made to comments after they are posted (a feature available on Reddit but not on Twitter).

Text Analysis Measures

The impact of Russian IRA activity on the conversational quality on Reddit was operationalised using three text analysis measures, which were applied to each post included in the analysis: Integrative Complexity, Toxicity and Identity Attack.

Integrative Complexity (IC) is a social-psychological measure of how much an individual presents the ability to think and reason with input from multiple perspectives.[72] Higher IC is associated with more accurate and balanced perceptions of other people, lower prejudice, the use of more information when making decisions, as well as less extreme views.[73] Lower IC in discussions is associated with prediction of future violence and intergroup conflict.[74] We used AUTO IC[75] to get IC scores for each Reddit post. The system produces a score from 1 to 7 for each comment, with lower scores representing lower levels of complexity. AUTO IC has been used successfully for the study of online terrorist content, demonstrating the validity of applying the measure to the digital domain.[76]

[71] Pushshift, Reddit.

[72] S. Streufert and P. Suedfeld, 'Conceptual Structure, Information Search, and Information Utilization', *Journal of Personality and Social Psychology* Vol. 2, № 5 (November 1965): 736–40.

[73] Allison Smith, Peter Suedfeld, Lucien G. Conway III, and David G. Winter, 'The Language of Violence: Distinguishing Terrorist from Nonterrorist Groups by Thematic Content Analysis', *Dynamics of Asymmetric Conflict* Vol. 1, Issue 2 (July 2008): 142–63; Philip E. Tetlock, Randall S. Peterson, and Jane M. Berry, 'Flattering and Unflattering Personality Portraits of Integratively Simple and Complex Managers', *Journal of Personality and Social Psychology* Vol. 64, № 3 (1993): 500–11.

[74] Karen Guttieri, Michael D. Wallace, and Peter Suedfeld, 'The Integrative Complexity of American Decision Makers in The Cuban Missile Crisis', *Journal of Conflict Resolution* Vol. 39, № 4 (1 December 1995): 595–621; Peter Suedfeld and Susan Bluck, 'Changes in Integrative Complexity Prior to Surprise Attacks', *Journal of Conflict Resolution* Vol. 32, № 4 (1988): 626–35.

[75] Lucien G. Conway III, Kathrene R. Conway, Laura Janelle Gornick, and Shannon C. Houck, 'Automated Integrative Complexity', *Political Psychology* Vol. 35, № 5 (2014): 603–24; Shannon C Houck, 'Automated Integrative Complexity : Current Challenges and Future Directions', *Political Psychology* 35, Issue 5 (2014): 603–24.

[76] Shannon C. Houck, Meredith A. Repke, and Lucian G. Conway III, 'Understanding What Makes Terrorist Groups' Propaganda Effective: An Integrative Complexity Analysis of ISIL and Al Qaeda', *Journal of Policing, Intelligence and Counter Terrorism* Vol. 12, Issue 2 (2017): 105–18.

We measured the level of Toxicity of each Reddit post with the Google Perspective API.[77] This classification tool was designed by Google's 'Project Jigsaw' and 'Counter Abuse Technology' teams with the aim of promoting better discussions online.[78] The tool uses machine learning models to score the perceived impact a comment might have on a conversation. Comments defined as being ruder, more disrespectful, or more unreasonable receive a higher Toxicity score. The model gives a Toxicity score for each comment on a scale ranging from 0 to 1.

The Google Perspective API also provides additional classifiers that are more specific and can provide further insight into the nature of comments. The Identity Attack option measures the degree to which a comment demonstrates negative or hateful comments targeting someone because of their identity. This is especially useful in the current study as it measures specific intergroup aggression and conflict based on who people are perceived to be. As with Toxicity, the model provides an Identity Attack score for each post on a scale ranging from 0 to 1.

Analysis of submissions and comments

Russian IRA activity consisted of submissions and comments. A *submission* is the starting post for a new conversation—i.e. threads started by Russian IRA accounts—while a *comment* is a post made on an existing conversation thread started by a genuine user. We analysed submissions and comments separately. We tested whether threads started by Russian IRA posts differed from those started by genuine users, and if Russian IRA comments injected into an existing thread had an impact on the subsequent conversation.

To measure the impact of submissions from Russian IRA accounts, we collected all comments made on threads started by Russian IRA accounts, including the initial submission starting the conversation, from the eleven subreddits identified above. In total this included 2,368 submissions and 30,112 comments. To test whether these conversations differed from genuine conversations, we collected a similar number of random 'control' submissions to the same subreddits within the same time frame. As with the Russian IRA submissions, we collected all the responses to this sample of genuine submissions, with a resulting total of 1,872

[77] Google Project Jigsaw, 'Perspective', accessed 23 March 2018.

[78] Ellery Wulczyn, Nithum Thain, and Lucas Dixon, 'Ex Machina: Personal Attacks Seen at Scale', *Proceedings of the 26th International Conference on World Wide Web*, 2017, 1391–99.

submissions and 22,503 comments. The lower number of genuine submissions is due to the exclusion of some submissions which received no subsequent comments. We then compared the conversation qualities for these two types of threads (those started by Russian IRA posts versus genuine submissions). As each subreddit was likely to include both types of conversation, we compared like-for-like conversations in each subreddit independently. For each comment in a thread we calculated a number of metrics relating to the measures used to determine the quality of the conversation, namely Integrative Complexity, Toxicity and Identity Attack.

To measure the impact of Russian IRA comments on existing genuine threads (rather than on new threads), we collected the comments from all threads that received at least one comment from a Russian IRA account. The sample of unmanipulated comment threads above was also used as the control for this sample. This dataset contained 455,300 comments from 826 threads, 1,253 of which came from Russian IRA controlled accounts. For each thread we numbered all comments in chronological order, with the injected Russian IRA post numbered as index position zero, subsequent posts incremented positively and previous posts negatively. We limited our analysis to threads containing ≥ 20 comments and to the 100 posts either side of the injected Russian IRA post. For each of these 200 comments we calculated the three text analysis measures and averaged these for each position in the thread across all threads, to show the average trend of the conversations. The data were then analysed using a causal analysis model (see details below) to detect changes in the three metrics after the injection of a Russian IRA comment. The analysis was performed across all subreddits for each metric. To explore whether the effect differed between political and non-political conversations, it was then run separately on political and non-political subreddits (Political_Subreddits; 'The_Donald', 'politics', 'Bad_Cop_No_Donut', 'PoliticalHumor', 'racism', 'news', 'worldnews', Other_Subreddits; 'aww', 'gifs', 'funny', 'AskReddit'). We investigated both the immediate and the overall impact of a content injection by running the analysis on the first 25 comments as well as on all 100 comments post-injection.

Statistical Methods

We investigated differences between conversations started by Russian IRA accounts compared to controls by using linear mixed models (LMMs) with the lme4 package. We investigated differences in Integrative Complexity, Toxicity, and Identity Attack between the Russian IRA-started and genuine threads, including subreddit ID as a random effect. Significance levels of fixed effects were obtained by comparing the full model to the null model with an χ^2 test. The difference between Russian IRA-started and genuine threads was also compared in each of the 11 individual subreddits using Welch two sample *t*-tests comparing the differences in mean conversation qualities. Toxicity and Identity Attack measures were square-root-transformed to ensure normality. Integrative Complexity could not be normalised, and so a Wilcoxon rank sum test with continuity correction was used. We corrected for multiple comparisons by adjusting the p-values with a Bonferroni-Holm correction.

We calculated the impact of a single artificial comment on an existing thread using the CausalImpact() package for R.[79] This package constructs a Bayesian structural time-series model to estimate the causal effect of a specific event on a time-series. In this case the time-series is the conversation quality (taken as the three text analysis measures) as it progresses over time along the thread, and the event is the Russian IRA comment injection at index position zero. This method allowed us to make causal inferences even though performing a randomised experiment was not possible. Through the construction of a time-series model this method predicts a counterfactual of how the response metric would have evolved after the intervention if the intervention had never occurred.[80] This method requires a control time-series of similar data unaffected by the intervention—here we used the unmanipulated threads. By calculating the relationship between the control and response time-series trends on the 100 posts prior to the intervention, the model then predicted the response time-series over the subsequent 100 posts, had there been no injection of Russian IRA comment. We then calculated the observed pointwise differences between manipulated and predicted threads after the intervention occurred. Summing these pointwise differences over a given time window, the model could provide a measure of the size of this cumulative difference over time, which was tested for statistical significance with a Bayesian one-sided tail area probability test.

[79] Kay H Brodersen and Alain Hauser, 'Package "CausalImpact"', *CRAN*, 2017, 1–8.

[80] Kay H. Brodersen, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L. Scott, 'Inferring Causal Impact Using Bayesian Structural Time-Series Models', *The Annals of Applied Statistics* Vol. 9, № 1 (2015): 247–74.

Results

Polarisation of Twitter conversations

Correlations between Russian IRA activity and subsequent polarisation of the Twitter conversation related to Black Lives Matter were significantly higher than expected by chance (permutation test, Figure 3b). This effect did not occur immediately following Russian IRA activity, but rather occurred predominantly between 3 and 10 days after the conversation manipulation had taken place. More specifically, it increased over time until reaching a peak around 7–9 days following the activity, and then gradually returned to the initial base level (Figure 3a). The effect started earlier, lasted longer, and was more pronounced when we considered Russian IRA activity over a longer time window (Figure 3, Table 3, see Table S1 for individual significance scores and correlation effect sizes). When looking at the longest period of cumulative activity—seven days—this trend appeared to last for almost two weeks from day two until day 14. Importantly, there was no general increasing or decreasing trend over time for either Russian IRA activity or polarisation and so our results were not due to long-term matching trends between the two variables.

The distributions of daily Russian IRA activity showed a long right tail (Appendix Figure 1c), suggesting this activity was uncommonly large on certain days. We tested whether these spikes in Russian IRA activity had an especially large effect on subsequent conversation polarisation by taking the top 10 days with the highest degree of polarisation, and testing whether each of these days had been preceded by a spike in Russian IRA activity (defined as a day with over 100 tweets) within a period of 10 days. We found that in 80% of these most polarised days, a spike in activity had preceded the polarisation.

The highest peaks in Russian IRA activity were fairly evenly distributed throughout the period studied. The mean Russian IRA activity across all days was 27 tweets, but this spiked as high as 592 tweets in a single day and 16 days had over 100 posts.

Reddit submissions

The conversation quality on threads started by Russian IRA-operated accounts differed substantially from that of genuine conversations, but the direction of these differences varied between subreddits and thus between topics of conversation. Overall, posts on threads started by Russian IRA accounts had higher Toxicity (IRA: 0.48 ± 0.001 vs genuine: 0.47± 0.002, n = 56,249, χ_1^2 =

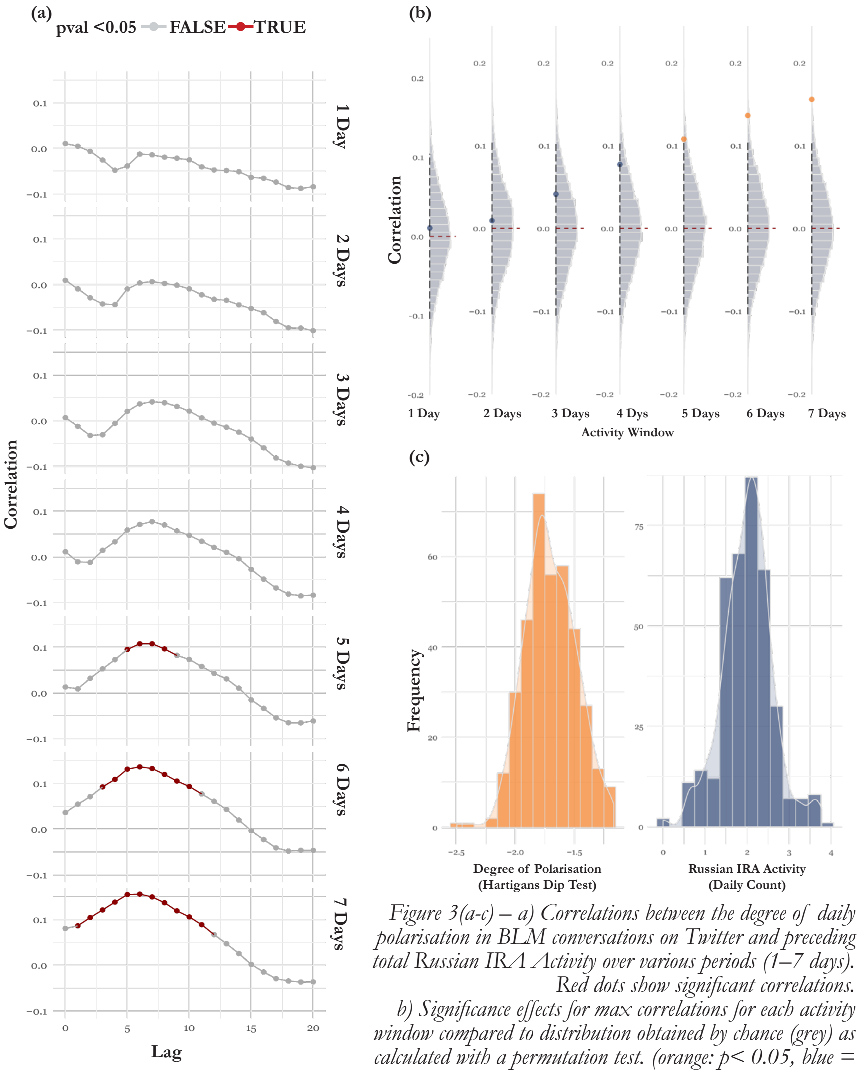

Figure 3(a-c) – a) Correlations between the degree of daily polarisation in BLM conversations on Twitter and preceding total Russian IRA Activity over various periods (1–7 days). Red dots show significant correlations.

b) Significance effects for max correlations for each activity window compared to distribution obtained by chance (grey) as calculated with a permutation test. (orange: p< 0.05, blue = non-significant)

c) Normalised distributions of polarisation and activity (see appendix for raw distributions)

	Number Significant Days	Start Day (Lag)	End Day (Lag)	Day of Max Correlation	Max Correlation	p
1	0	NA	NA	1	0.011	0.419
2	0	NA	NA	1	0.009	0.428
3	0	NA	NA	7	0.041	0.217
4	0	NA	NA	7	0.078	0.069
5	4	5	8	7	0.107	0.019
6	8	3	10	6	0.136	0.005
7	11	2	11	6	0.156	<0.001

Table 3 – Statistical results for the highest correlation in the lagged permutated test across each activity window. For full results see annexe Table 1.

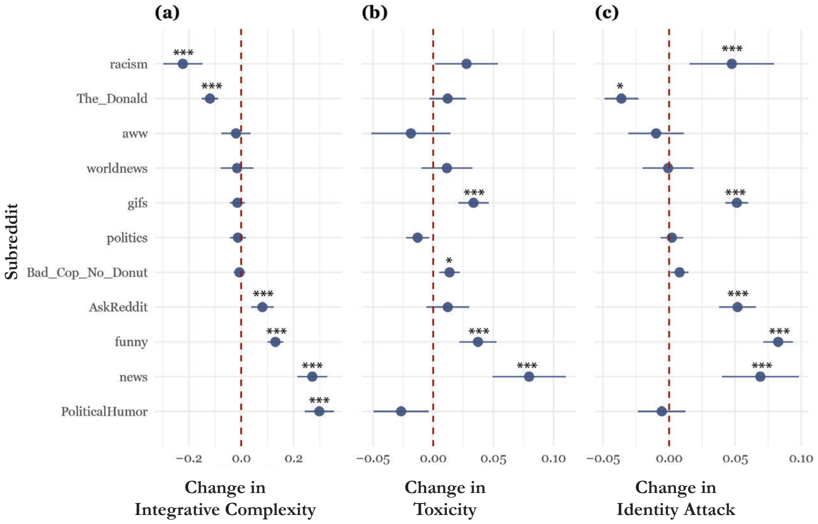

Figure 4(a-c) – Differences in mean conversation quality scores for threads started by Russian IRA Reddit accounts compared to genuine comment threads within the same subreddit. Higher values indicate Russian IRA started conversations scored higher on that conversation metric.

28.34, $p < 0.001$) and Identity Attacks (IRA: 0.42 ± 0.001 vs genuine 0.40± 0.001, $\chi_1^2 = 85.33$, $p < 0.001$) but showed no overall change in Integrative Complexity (IRA: 1.37± 0.004 vs genuine 1.36 ± 0.004, $\chi_1^2 = 2.39$, $p = 0.122$).

Further analyses performed on individual subreddits showed that threads started by Russian accounts within r/news, r/gifs, r/funny and r/Bad_Cop_No_Donut had higher average Toxicity scores than genuine threads in the same subreddits (Figure 4b, Table 4). Other subreddits showed no differences. We found a similar pattern with regard to levels of Identity Attack. Threads started by Russian accounts within r/racism, r/news, r/gifs, r/funny and r/AskReddit had higher average Identity Attack scores than genuine threads in the same subreddits (Figure 4c, Table 4), while artificial comment threads started within r/TheDonald by comparison had a lower average Identity Attack scores than genuine threads. Other subreddits showed no differences. While we found no difference in Integrative Complexity overall, artificial threads started by Russian IRA accounts received comments with lower IC scores in r/TheDonald and r/racism, but higher IC scores in r/PoliticalHumor, r/news, r/funny and r/AskReddit (Figure 4a, Table 4).

Text Analysis Measures

Subreddits	Integrative Complexity					Toxicity					
	Mean IRA Started	Mean Genuine	W	p	d	Mean IRA Started	Mean Genuine	df	t	p	d
r/racism	1.18±0.02	1.40±0.03	132938	< 0.001	0.318	0.55±0.01	0.52±0.01	987	-2.09	0.037	0.127
r/The_Donald	1.15±0.01	1.27±0.01	1626926	< 0.001	0.149	0.48±0.01	0.47±0.01	3646	1.56	0.471	0.051
r/aww	1.09±0.02	1.11±0.02	32751	1	0.064	0.36±0.01	0.38±0.01	292	-1.11	0.541	0.103
r/worldnews	1.41±0.03	1.43±0.02	426810	1	0.024	0.46±0.01	0.45±0.01	1023	1.06	0.541	0.033
r/gifs	1.22±0.01	1.23±.01	3296930	1	0.029	0.45±0.004	0.42±0.01	4900	-5.22	< 0.001	0.145
r/politics	1.44±0.01	1.45±0.01	9524468	1	0.018	0.45±0.004	0.46±0.002	4083	2.66	0.06	0.06
r/Bad_Cop_No_Donut	1.38±0.01	1.39±0.01	18150512	1	0.011	0.53±0.003	0.51±0.003	11,717	-3.16	0.013	0.058
r/AskReddit	1.34±0.02	1.26±0.01	858343	< 0.001	0.149	0.43±0.01	0.42±0.01	2413	1.34	0.541	0.054
r/funny	1.29±0.01	1.16±0.01	1690132	< 0.001	0.259	0.46±0.004	0.42±0.001	2143	4.75	< 0.001	0.164
r/news	1.42±0.001	1.15±0.03	1296950	< 0.001	0.406	0.49±0.002	0.41±0.02	186	-5.18	< 0.001	0.334
r/PoliticalHumor	1.53±0.02	1.23±0.02	664965	< 0.001	0.405	0.44±0.002	0.41±0.01	726	4.75	0.136	0.164

Table 4 – Statistical results for pared sample t-tests comparing differences in mean conversation quality scores for threads started by Russian IRA Reddit accounts compared to organic comment threads within the same subreddit

Identity Attack

Subreddits	Mean IRA Started	Mean Genuine	df	t	p	d
r/racism	0.61±0.01	0.56±0.01	984	-2.92	0.022	0.177
r/The_Donald	0.39±0.01	0.43±0.01	3792	5.5	< 0.001	0.178
r/aww	0.29±0.01	0.30±0.01	294	-0.92	1	0.085
r/worldnews	0.42±0.01	0.42±0.01	991	-0.07	1	0.004
r/gifs	0.36±0.003	0.31±0.003	5208	-11.7	< 0.001	0.314
r/politics	0.42±0.003	0.42±0.002	3766	0.5	1	0.012
r/Bad_Cop_No_Donut	0.43±0.003	0.42±0.003	11,644	-2.35	0.095	0.043
r/AskReddit	0.37±0.01	0.32±0.004	2112	-7.31	< 0.001	0.3
r/funny	0.40±0.004	0.32±0.004	3155	-14.6	< 0.001	0.429
r/news	0.44±0.002	0.34±0.02	185	-4.7	< 0.001	0.314
r/PoliticalHumor	0.39±0.003	0.40±0.01	735	-0.6	1	0.031

Table 4 – Statistical results for pared sample t-tests comparing differences in mean conversation quality scores for threads started by Russian IRA Reddit accounts compared to organic comment threads within the same subreddit (continued)

Results – Reddit comments

Across all subreddits and comment threads, Russian IRA comments led to a small drop in the Integrative Complexity of the subsequent conversation over a period of 100 comments by a factor of 1% ± 0.51 (Figure 5a-c).

For the period after a Russian IRA comment injection, the average Integrative Complexity was 1.41 ± 0.004. In the absence of any intervention, the causal analysis model predicted an average value of 1.42 ± 0.006, significantly higher than the observed value (Bayesian one-sided tail area probability $p = 0.035$). In other words, on average a Russian comment caused a 0.01 decrease in IC compared to predictions.

Additionally, Russian IRA comment injections lead to short term increase in the Integrative Complexity of conversations in non-political subreddits by a factor of 2% ± 0.77 over the subsequent 25 comments ($p = 0.005$).

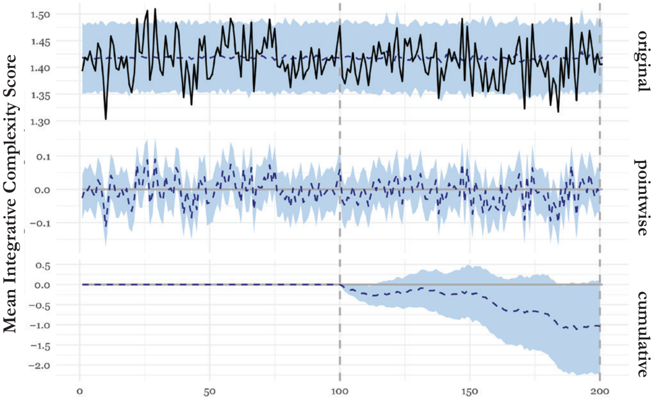

Figure 5(a-c) – Causal Impact analysis of Artificial Russian Reddit account comment injection on the Integrative Complexity of the conversation. Panel (a) - the observed trend for average IC over the course of the conversations, along with the counterfactual prediction period after the intervention if it had not occurred. Panel (b) - the pointwise difference this counterfactual prediction and the observed data. Panel (c) - the cumulative pointwise difference overtime, giving an indication of the overall effect of the intervention on the IC of the conversation.

There were no measurable differences in the effect of Russian IRA comment injection on Integrative Complexity in political subreddits when considered in isolation, or in non-political subreddits over longer periods of time (Table 5).

			Text Analysis Measures											
			Integrative Complexity				**Toxicity**				**Identity Attack**			
		Time Span (Comments)	**Mean Observed**	**Mean Predicted**	**P**	**Cumulative Change**	**Mean Observed**	**Mean Predicted**	**P**	**Cumulative Change**	**Mean Observed**	**Mean Predicted**	**P**	**Cumulative Change**
Subreddits	**All**	**100**	**1.41±0.004**	**1.42±0.006**	**0.035**	**-1%±0.51**	0.26±0.004	0.27±0.001	0.148	-1%±0.77	0.20±0.001	0.20±0.001	0.064	-1%±0.51
		25	1.41±0.003	1.42±0.002	0.152	-1%±0.77	0.27±0.0005	0.27±0.002	0.303	0%±0.77	0.20±0.001	0.21±0.002	0.379	0%±1.02
	Political	**100**	1.43±0.006	01.43±.01	0.373	0%±0.77	0.27±0.0003	0.27±0.002	0.301	1%±1.28	0.21±0.0002	0.21±0.002	0.085	2%±1.28
		25	1.42±0.02	1.43±0.03	0.178	-1±1.02	**0.28±0.001**	**0.27±0.002**	**0.019**	**3%±1.53**	0.21±0.001	0.21±0.004	0.137	2%±2.6
	Non-Political	**100**	1.23±0.005	1.24±0.007	0.329	0%±0.77	0.23±0.003	0.23±0.002	0.378	0%±1.28	0.13±0.0002	0.13±0.002	0.482	0%±1.53
		25	**1.26±0.007**	**1.23±0.002**	**0.005**	**2%±0.77**	0.22±0.001	0.23±0.004	0.118	-2%±1.80	**0.14±0.001**	**0.13±0.002**	**0.001**	**10%±2.04**

Table 5 – Statistical results for causal impact analysis across the three conversation measures; Integrative Complexity, Toxicity and Identity Attack and across two time periods; 100 and 25 comments

Russian IRA comment injections also affected the Toxicity of subsequent conversations, but these effects occurred only in political subreddits and for short periods. While there was no significant effect of Russian IRA comment injection on Toxicity if considered over the entire post-intervention period of 100 comments, comment injections did increase Toxicity of the conversation over the next 25 comments by a factor of 3% ± 1.53 (p = 0.019), but this effect subsequently disappeared over the following 75 comments (Figure 6).

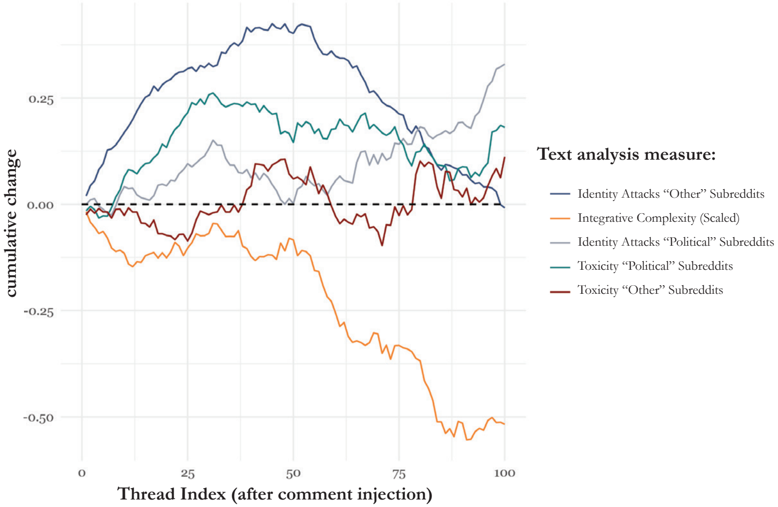

Figure 6 – Cumulative impact of artificial Russian IRA comment injections on the Toxicity, Identity Attack (IA), and Integrative Complexity (IC) of subsequent conversation on Reddit in political and non-political subreddits.

Similarly, the impact of the degree of Identity Attacks taking place in conversations after a Russian comment injection also depended on whether the comments occurred in political or non-political subreddits. In non-political subreddits, comment injection was followed by a marked short-term increase in Identity Attacks over the next 25 comments by a factor of 10% ± 2.04 (p = 0.001), and this effect subsequently dissipated over time. There was no change in the degree of Identity Attacks following a Russian IRA comment injection in a political subreddit.

Discussion

In this study we examined whether social media activity from artificial accounts run by the Russian IRA led to measurable changes in the conversation of genuine users on Twitter and Reddit. Our results show that Russian IRA activity indeed predicted changes in the conversations taking place on both platforms, but the exact effects differed between platforms and the type of manipulation taking place.

On Twitter, higher amounts of Russian IRA activity in the Black Lives Matter conversation predicted increases in the subsequent conversational polarisation of genuine Twitter users. This increase in polarisation peaked approximately one week after the injection of Russian IRA content and the association was most pronounced around the periods of highest Russian activity, suggesting that large spikes in Russian IRA activity had the greatest influence on the subsequent conversation. The gradual build-up of these effects over a week may reflect a structural property of Twitter—that more a tweet is retweeted, the more influence it gains on the network.[81] On days with higher numbers of tweets from Russian IRA accounts there was a greater likelihood that one of the tweets would go 'viral' and be exposed to a much larger audience—either by simply manually increasing the number of tweets or by mass (automated) retweeting through the use of connected bot accounts.[82] Earlier research has found that Russian IRA accounts embed themselves into both for and against sides of the Black Lives Matter debate;[83] our results show that this may have contributed to the polarisation of both sides of the debate. It is noteworthy that we find this effect despite the high attrition rate within our Twitter data; 45% of Tweets were deleted before data collection. Deleted tweets are more likely to contain negative sentiment or profanity[84] or to be 'regretted' by their author,[85] and so the exclusion of these tweets likely muted the observed effects of Russian IRA activity on polarisation.

On Reddit we found that threads started by Russian IRA accounts were generally more Toxic than conversations started by genuine users whilst also showing more instances of Identity Attacks. Higher Toxicity reflects that these conversations

[81] Ee-Peng Lim, Palakorn Achananuparp, and Feida Zhu, 'On Modeling Virality of Twitter Content', *ICADL*, 2011, 212:221.
[82] Kumar et al., An Army of Me'; Fredheim, 'Robotrolling'.
[83] Arif et al ., 'Acting the Part'.
[84] Parantapa Bhattacharya and Niloy Ganguly, 'Characterizing Deleted Tweets and Their Authors', *Icwsm*, 2016, 10–13.
[85] Lu Zhou, Wenbo Wang, and Keke Chen, 'Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones', *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 2016, 603–12.

were more rude, aggressive, or disrespectful, and more likely to inflame other users (both targets and observers), while conversations with higher Identity Attack scores contained a greater number of hostile comments made against people due to group membership, including race and political affiliation.[86] Both of these measures indicate that Russian IRA activity was effective in promoting hostile conversations among other users, likely increasing divisions among group lines.

The effect of Russian IRA activity on Integrative Complexity was more complicated. While there was no overall difference in the Integrative Complexity of threads started by the Russian IRA compared to genuine threads, there were differential effects of Integrative Complexity depending on the subreddit in which a conversation was started. Conversations started by Russian IRA accounts in r/racism and r/The_Donald showed reductions in Integrative Complexity, (less complex conversations with less nuance, demonstrating reasoning from fewer viewpoints)[87], while conversations started in r/AskReddit, r/funny, r/news and r/PoliticalHumor displayed higher Integrative Complexity compared to genuine conversation threads in these subreddits. One interpretation of these results is that they are related to the partisan nature of the political subreddits, which may facilitate a reduction in complexity due to a lack of opposing voices,[88] compared to the 'general interest' subreddits, which may enable greater intergroup discussion because of their non-partisan nature. These and other explanations need direct testing, however, and merit further research.

We also found evidence suggesting a causal relationship between Russian IRA activity and conversation quality by studying the impact of comments from Russian IRA accounts injected into existing genuine conversations. Across all subreddits, Russian IRA comment injections led to a decrease in the Integrative Complexity of the conversation over the subsequent 100 comments. Additionally, there was a shorter-lived effect, detectable on the 25 subsequent comments, which led to an increase in Toxicity in political subreddits and an increase in the level of Identity Attacks in non-political subreddits. Although these findings are less clear-cut than those described above, they similarly demonstrate that any measurable effects of Russian IRA activity are in the direction of undermining conversational quality. Cumulatively, these small effects have the power significantly to shape a conversation. They also suggest that different dynamics unfold in political and

[86] Google Project Jigsaw, 'Perspective'; Wulczyn et al., 'Ex Machina'
[87] Streufert and Suedfeld, 'Conceptual Structure'.
[88] Sunstein, *#Republic*; Pariser, *The Filtter Bubble.*

non-political online conversations, which is in line with previous findings,[89] and that distinguishing between these conversational domains remains important in future research. We found that in the absence of manipulation, the control threads within political subreddits had higher Integrative Complexity, Toxicity and Identity Attacks than non-political subreddits, suggesting that political conversations are characterised by both increased engagement and increased hostility. These characteristics may relate to findings that echo chambers form primarily in the political domain,[90] but whether these are causes or effects remains to be tested.

Comparing the results across platforms, we found that the effects of Russian IRA activity manifested more quickly on Reddit than on Twitter. On average, the effects detected over 25 and 100 Reddit posts following manipulation peaked around 3.5 days and 5 days after submission respectively, while on Twitter the association between Russian IRA activity and polarisation peaked after 7 days. This is likely due to the structural differences between the platforms. On Twitter the impact of content is measured by popularity—how many people react to it—and therefore tweets that go viral can have a large effect on the overall conversation.[91] On Reddit a single comment cannot go viral and impact results from the cumulative effect of many posts or of many users 'upvoting' a thread.[92] On Twitter, tweets can take longer to go viral, compared to the direct responses which occur on Reddit threads, that have a shorter-lived visibility. Given these considerations, it would also be interesting to study the consequences of more sustained periods of Russian activity in a single Reddit thread. Our analytical procedure did not allow us to identify these consequences as we could only model a single intervention at a time, but we expect that repeated co-ordinated activity within a single thread would lead to increased cumulative effects.[93] Including this co-ordinated behaviour may mean that the consequences of comments in existing threads more closely resemble the observed differences in total conversations following genuine submissions and Russian IRA submissions.

[89] Barberá et al., 'Tweeting From Left to Right'; Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis, 'Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship', *International World Wide Web Conference* 2 (2018).
[90] Barberá et al., 'Tweeting From Left to Right'; Garimella et al., 'Political Discourse on Social Media'.
[91] Ee-Peng Lim, Palakorn Achananuparp, and Feida Zhu, 'On Modeling Virality of Twitter Content', *ICADL*, 2011, 212:221.
[92] Amir Salihefendic, 'How Reddit Ranking Algorithms Work', Hacking and Gonzo, *Medium*, 8 December 2015.
[93] J. M. Berger and Jonathon Morgan, 'The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter', *The Brookings Project on U.S. Relations with the Islamic World*, 5 March 2015; Emilio Ferrara, 'Manipulation and Abuse on Social Media', *SIGWEB Newsletter*, Spring 2015.

By increasing the polarisation of conversations on Twitter and undermining the quality of conversations on Reddit, Russian IRA activity is likely to be effective in increasing the distance between social groups, fuelling both ideological and affective polarisation.[94] This in turn provides ideal circumstances for the distribution of disinformation[95] because it increases the acceptance of (inaccurate) information that confirms prior views—a phenomenon known as 'confirmation bias'[96]—and facilitates repeated exposure to the same inaccurate information because alternative perspectives are eliminated from discussion by design.[97] Western societies that focus more on internal strife from polarised domestic communities tend to focus less on international issues, illustrating that this activity may be part of a larger geopolitical strategy.[98]

In this study we focused on activity originating from publicly attributed Russian IRA accounts and their effect on two key social media platforms. Future research should consider including other platforms, and also other groups engaged in information operations. Russian IRA activity accounts for a fraction of all possible information operations activities worldwide, and many other groups produce similar content for a range of different purposes. This includes pursuing international strategic goals (as demonstrated by Iranian actions),[99] focusing attention on perceived domestic concerns (utilised by far-right groups),[100] and quashing dissent (a tactic favoured by China).[101] Our study only begins to unveil the negative effect of information operations globally. If fuelling arguments on both sides of controversial topics works to increase polarisation in these conversations, then pushing only a single side may work to decrease polarisation or even to stifle active debate. This might be the goal for a regime that wishes to quash dissent or opposition. For example, evidence of Chinese government involvement in online discussions shows that across ~450 million social media posts per year the strategy is not to engage with controversial topics or with sceptics of government, but rather to change the subject

[94] Mason, 'I Disrespectfully Agree'.
[95] Garrett et al., 'Driving a Wedge Between Evidence and Beliefs'; Michela Del Vicario, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo, 'Public Discourse and News Consumption on Online Social Media: A Quantitative, Cross-Platform Analysis of the Italian Referendum', arXiv.org, February 2017.
[96] Raymond S Nickerson, 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises', *Review of General Psychology* Vol. 2, № 2 (1998): 175–220.
[97] Pennycook et al., 'Prior Exposure Increases'; Berinsky, 'Rumors and Health Care Reform'.
[98] P. W. Singer and Emerson T. Brooking, *Likewar: The Weaponization of Social Media* (New York: Houghton Mifflin Harcourt Publishing Company, 2018); James Kirchick, 'Russia's Plot against the West', *Politico*, 17 March 2017.
[99] Karan et al., '#TrollTracker'.
[100] Nathaniel Gleicher, 'Removing Coordinated Inauthentic Behavior from the UK and Romania', Facebook Newsroom, 7 March 2019.
[101] Gary King, Pan Jennifer, and Roberts Margaret E, 'How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument', *American Political Science Review* Vol. 111, Issue 03 (2017): 484–501.

of conversations with vocal cheerleading for pro-China positions to overwhelm opposition voices.[102] The Kremlin takes a similar approach towards domestic audiences, using troll farms such as the Russian IRA to produce vast quantities of pro-regime messages in Russian for local consumption.[103]

While it remains to be seen whether these online effects translate into offline actions, there is evidence that online activities can have substantial effects on real world behaviour ranging from exercise and smoking to consumer trends.[104] Our research also shows that online interaction between groups predicts offline violence,[105] while other research demonstrates how online aggression towards disadvantaged groups can lead to offline hate crimes.[106] By demonstrating that information operations promote social polarisation and can have measurable impacts on online conversations more broadly, our study also highlights the risk of potential future vulnerabilities. The ability of hostile actors to create polarising content is increasing at pace, thanks to advances in machine-generated text that closely resembles human speech.[107] If this technology is paired with malicious intent to drive communities apart using social media platforms, then the volume of content may well expand and increase the severity of the challenge to detecting inauthentic content and oppose it.[108]

It is therefore essential to design solutions that address and counter the negative effects of hostile information operations. Identifying the impact of information operations is only the first step in creating counter measures. Evidence suggests that organised attempts to challenge the veracity of disinformation on Twitter

[102] King, Jennifer, and Margaret E. 'How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument' *American Political Science Review*, 2017, 111, 3, 484–501

[103] Gallacher and Fredheim, 'Division Abroad, Cohesion at Home'.

[104] Tim Althoff, Pranav Jindal, and Jure Leskovec, 'Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behaviour', *WSDM Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, 537–46; Jacob B. Depue, Brian G. Southwell, Anne E. Betzner, and Barbara M. Walsh, 'Encoded Exposure to Tobacco Use in Social Media Predicts Subsequent Smoking Behavior', *American Journal of Health Promotion* Vol. 29, № 4 (2015): 259–61; Sidharth Muralidharan and Linjuan Rita Men, 'How Peer Communication and Engagement Motivations Influence Social Media Shopping Behavior: Evidence from China and the United States', *Cyberpsychology, Behavior, and Social Networking* Vol. 18, № 10 (2015): 595–601.

[105] John David Gallacher, Marc W Heerdink, and Miles Hewstone, 'Online Contact between Opposing Political Protest Groups via Social Media Predicts Physical Violence of Offline Encounters', (*under review*), 1–44.

[106] Karsten Müller and Carlo Schwarz, 'Fanning the Flames of Hate: Social Media and Hate Crime', *SSRN Electronic Journal*, 7 December 2017.

[107] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, 'Language Models Are Unsupervised Multitask Learners', 2018.

[108] Miles Brundage et al., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', February 2018

are generally ineffective,[109] while spontaneous fact-checking on Facebook is rare and generally unsuccessful.[110] Other technical solutions should therefore focus on the early detection of artificial content before it can manipulate online conversations,[111] or educational methods which may mitigate the effects of disinformation through inoculation of citizens.[112] Structural changes to social media platforms promoting positive exposure to members of opposing groups will also likely reduce and dilute the impact of efforts to divide these same groups through negative content injections.[113] Addressing the challenge of disinformation is so broad that designing effective interventions will require interdisciplinary efforts at multiple levels of analysis.[114]

Conclusion

Our study reveals that the malicious use of social media by 'fake' accounts can measurably affect the subsequent conversations held by genuine users. Using the activity of the Russian Internet Research Agency on Twitter and Reddit as case studies, we have shown that this effect differed between social media platforms. The effect of Russian activity on Twitter was to increase polarisation after a one-week delay, while there was a more immediate effect on Reddit, immediately altering the quality of subsequent conversations. By developing methods to measure the impact of information operations in online conversations, our study provides an important step in developing effective countermeasures.

Acknowledgements

This research was supported by grants from EPSRC and the University College Oxford Radcliffe Scholarship. The second author's contribution to this project was partially supported by a grant from the Netherlands Organisation for

[109] Jieun Shin, , Lian Jian, Kevin Driscoll, and François Bar, 'Political Rumoring on Twitter during the 2012 US Presidential Election: Rumor Diffusion and Correction', *New Media and Society* Vol. 19, № 8 (2017): 1214–35; Drew B. Margolin, Aniko Hannak, and Ingmar Weber, 'Political Fact-Checking on Twitter: When Do Corrections Have an Effect?', *Political Communication* Vol. 35, 2 (2018): 196–219.

[110] Adrien Friggeri, Lada A. Adamic, Dean Eckles, and Justin Cheng, 'Rumor Cascades', *ICWSM 2014 International Conference on Weblogs and Social Media*, 2014, 101–10.

[111] Jordan Wright and Olabode Anise, 'Don't @ Me : Hunting Twitter Bots at Scale', *Black Hat*, 2018, 1–43.

[112] Jon Roozenbeek and Sander Van Der Linden, 'The Fake News Game: Actively Inoculating Against the Risk of Misinformation', *Journal of Risk Research* 9877 (2018): 1–11.

[113] Rupert Brown and Miles Hewstone, 'An Integrative Theory of Intergroup Contact', *Advances in Experimental Social Psychology* Vol. 37 (2005): 255–343; Thomas F Pettigrew and Linda R Tropp, 'How Does Intergroup Contact Reduce Prejudice? Meta-Analytic Tests of Three Mediators', *European Journal of Social Psychology* Vol. 38, Issue 6 (2008): 922–34. the 21st century has not begun auspiciously. As we enter only its fifth year, we have already seen unprecedented incidents of international conflict and terrorism (e.g., Afghanistan, 2002; Iraq, 2003; Spain, 2004; USA, 2001

[114] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer et al., 'The Science of Fake News', *Science* Vol. 359, Issue 6380 (2018): 1094–96.

Scientific Research (NWO 446-16-015). The funding bodies played no further role in designing or implementing the research, and the authors declare no competing interests. We thank members of the Oxford Centre for the Study of Intergroup Conflict for their helpful feedback.

Appendix

Sum Period (Days)	Lag Period (Days)	Correlation	p
1	0	0.011	0.419
	1	0.005	0.467
	2	-0.006	0.548
	3	-0.026	0.688
	4	-0.048	0.820
	5	-0.038	0.768
	6	-0.013	0.599
	7	-0.014	0.610
	8	-0.019	0.648
	9	-0.022	0.665
	10	-0.025	0.690
	11	-0.040	0.784
	12	-0.047	0.821
	13	-0.049	0.822
	14	-0.051	0.832
	15	-0.063	0.887
	16	-0.065	0.894
	17	-0.074	0.920
	18	-0.086	0.951
	19	-0.088	0.955
	20	-0.084	0.945
2	0	0.009	0.428
	1	-0.009	0.564
	2	-0.029	0.709
	3	-0.043	0.788
	4	-0.044	0.798
	5	-0.010	0.573
	6	0.003	0.473
	7	0.006	0.451
	8	0.002	0.484
	9	-0.001	0.514
	10	-0.009	0.575
	11	-0.022	0.663
	12	-0.032	0.728
	13	-0.034	0.740
	14	-0.045	0.803
	15	-0.053	0.841
	16	-0.062	0.878
	17	-0.081	0.934
	18	-0.095	0.962
	19	-0.095	0.964
	20	-0.101	0.971
3	0	0.007	0.441
	1	-0.012	0.594
	2	-0.032	0.739
	3	-0.031	0.723
	4	-0.006	0.544
	5	0.020	0.354
	6	0.037	0.242
	7	0.041	0.217
	8	0.039	0.227
	9	0.031	0.280
	10	0.021	0.348
	11	0.006	0.459
	12	-0.005	0.541
	13	-0.014	0.611
	14	-0.025	0.680
	15	-0.040	0.772
	16	-0.059	0.864
	17	-0.082	0.936
	18	-0.093	0.960
	19	-0.100	0.969
	20	-0.103	0.971
4	0	0.011	0.409
	1	-0.011	0.581
	2	-0.012	0.590
	3	0.014	0.394
	4	0.033	0.264
	5	0.059	0.131
	6	0.071	0.087
	7	0.077	0.069
	8	0.070	0.092
	9	0.057	0.141
	10	0.047	0.184
	11	0.034	0.259
	12	0.020	0.345
	13	0.010	0.416
	14	-0.004	0.527
	15	-0.028	0.691
	16	-0.049	0.816
	17	-0.068	0.893
	18	-0.082	0.930
	19	-0.085	0.940
	20	-0.084	0.940
5	0	0.013	0.405
	1	0.009	0.438
	2	0.032	0.279
	3	0.053	0.156
	4	0.073	0.080
	5	0.095	0.035
	6	0.108	0.020
	7	0.108	0.020
	8	0.096	0.033
	9	0.082	0.058
	10	0.073	0.082
	11	0.058	0.136
	12	0.043	0.209
	13	0.031	0.283
	14	0.010	0.422
	15	-0.015	0.613
	16	-0.034	0.741
	17	-0.054	0.848
	18	-0.065	0.890
	19	-0.065	0.894
	20	-0.061	0.877
6	0	0.036	0.248
	1	0.054	0.151
	2	0.071	0.087
	3	0.092	0.040
	4	0.109	0.018
	5	0.131	0.006
	6	0.136	0.005
	7	0.133	0.006
	8	0.120	0.010
	9	0.105	0.023
	10	0.093	0.037
	11	0.077	0.073
	12	0.061	0.125
	13	0.043	0.208
	14	0.020	0.360
	15	-0.003	0.534
	16	-0.023	0.674
	17	-0.041	0.785
	18	-0.048	0.823
	19	-0.046	0.814
	20	-0.047	0.814
7	0	0.081	0.058
	1	0.086	0.049
	2	0.104	0.023
	3	0.121	0.009
	4	0.138	0.003
	5	0.155	0.001
	6	0.156	< 0.001
	7	0.150	0.002
	8	0.137	0.004
	9	0.119	0.012
	10	0.106	0.022
	11	0.089	0.046
	12	0.067	0.101
	13	0.047	0.185
	14	0.026	0.313
	15	0.002	0.489
	16	-0.015	0.611
	17	-0.029	0.705
	18	-0.035	0.734
	19	-0.037	0.749
	20	-0.036	0.747

Table Appendix 1 – Statistical results for the lagged permutation test across activity window and lag period. Bold indicates statistical significance at the p = 0.005 level

DOI 10.30966/2018.RIGA.6.5.

Bibliography

Althoff, Tim, Pranav Jindal, and Jure Leskovec. 'Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behaviour', *WSDM Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, 2017, 537–46.

Arif, Ahmer, Leo G Stewart, and Kate Starbird. Acting the Part: Examining Information Operations within #BlackLivesMatter Discourse', *Proceedings of the ACM on Human-Computer Interaction* Vol. 2, Issue CSCW, Article № 20 (2018): 1–26.

Arnaudo, Dan. 'Computational Propaganda in Brazil: Social Bots during Elections', University of Oxford Computational Propaganda Research Project 8 (2017): 1–39.

Bail, Christopher, Lisa Argyle, Taylor Brown, John Bumpus, Haohan Chen, M.B. Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 'Exposure to Opposing Views Can Increase Political Polarization: Evidence from a Large-Scale Field Experiment on Social Media', *Proceedings of the National Academy of Sciences*, 2018, 1–6.

Bakshy, E., S. Messing, and L. A. Adamic. 'Exposure to Ideologically Diverse News and Opinion on Facebook', *Science* Vol. 348, Issue 6239 (2015): 1130–32.

Barberá, Pablo. 'Explaining the Spread of Misinformation on Social Media: Evidence from the 2016 U.S. Presidential Election', *Comparative Politics Newsletter* Vol. 28, Issue 2 (2018): 7–11.

Barberá, Pablo, John T Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting From Left to Right: Is Online Political Communication More than an Echo Chamber?', *Psychological Science* Vol. 26, № 10 (2015): 1531–42.

Barberá, Pablo, and Gonzalo Rivero. 'Understanding the Political Representativeness of Twitter Users', *Social Science Computer Review* Vol. 33, № 6 (2015), 712–29.

Bay, Sebastian, Giorgio Bertolin, Nora Biteniece, Edward H Christie, E Rolf, John D Gallacher, Kateryna Kononova, and Tetiana Marchenko. *Responding to Cognitive Security Challenges*. Edited by Anna Reynolds and Mike Collier. Riga, Latvia: NATO Strategic Communications Centre of Excellence.

Berger, J. M., and Jonathon Morgan. 'The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter', *The Brookings Project on U.S. Relations with the Islamic World*, 5 March 2015.

Berinsky, Adam J. 'Rumors and Health Care Reform: Experiments in Political Misinformation', *British Journal of Political Science* Vol. 47, Issue 2 (2017): 241–62.

Bhattacharya, Parantapa, and Niloy Ganguly. 'Characterizing Deleted Tweets and Their Authors', *Icwsm*, 2016, 10–13.

Bishara, Anthony J., and James B. Hittner. 'Testing the Significance of a Correlation with Nonnormal Data: Comparison of Pearson, Spearman, Transformation, and Resampling Approaches', *Psychological Methods* Vol. 17, № 3 (2012): 399–417.

Box, G. E. P., and D. R. Cox. 'An Analysis of Transformations', *Journal of the Royal Statistical Society* Vol. 26, № 2 (1964): 211–43.

Brady, William J., Julian A. Wills, John T. Jost, Joshua A. Tucker, and Jay J. Van Bavel. 'Emotion Shapes the Diffusion of Moralized Content in Social Networks', *Proceedings of the National Academy of Sciences* Vol. 114, № 28 (2017): 7313–18.

Brian, Ripley, Bill Venables, Douglas M Bates, David Firth, and Maintainer Brian Ripley. 'Package "MASS"' *CRAN* (2018).

Bright, Jonathan. 'Explaining the Emergence of Echo Chambers on Social Media: The Role of Ideology and Extremism', *SSRN Electronic Journal* (2016).

Brodersen, Kay H, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. 'Inferring Causal Impact Using Bayesian Structural Time-Series Models', *The Annals of Applied Statistics* Vol. 9, № 1 (2015): 247–74.

Brodersen, Kay H, and Alain Hauser. 'Package "CausalImpact"', *CRAN*, 2017, 1–8.

Brown, Rupert, and Miles Hewstone. 'An Integrative Theory of Intergroup Contact', *Advances in Experimental Social Psychology* Vol. 37 (2005): 255–343.

Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, et al. 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', February 2018.

Castree, Noel, Rob. Kitchen, and Alisdair. Rogers. 2013. *A Dictionary of Human Geography*. Oxford University Press.

DOI 10.30966/2018.RIGA.6.5.

Conover, M, J Ratkiewicz, and M Francisco. 'Political Polarization on Twitter', *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, Conference Paper (2011).

Crockett, M. J. 'Moral Outrage in the Digital Age', *Nature Human Behaviour* Vol. 1 (2017):769–71.

Depue, Jacob B., Brian G. Southwell, Anne E. Betzner, and Barbara M. Walsh. 'Encoded Exposure to Tobacco Use in Social Media Predicts Subsequent Smoking Behavior', *American Journal of Health Promotion* Vol. 29, № 4 (2015).

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson. The Tactics & Tropes of the Internet Research Agency', New Knowledge Disinformation Report Whitepaper, 2018.

Druckman, James N., Erik Peterson, and Rune Slothuus. 'How Elite Partisan Polarization Affects Public Opinion Formation', *American Political Science Review* Vol. 107, Issue 01 (2013).

Ferrara, Emilio. 'Manipulation and Abuse on Social Media', *SIGWEB Newsletter*, Spring 2015.

Ferrara, Emilio, Disinformation and social bot operations in the run up to the 2017 French presidential election', *First Monday*, Vol. 22, № 8, 2017.

Fredheim, Rolf. 2019. 'Robotrolling 2019, Issue 1' (Riga, Latvia, NATO StratCom COE, 2019).

Freelon, Deen, Charlton D. McIlwain, and Meredith Clark. 'Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice', *SSRN Electronic Journal*, (2016).

Friggeri, Adrien, Lada Adamic, Dean Eckles, and Justin Cheng. 'Rumor Cascades', *ICWSM 2014 International Conference on Weblogs and Social Media*, 2014, 101–10.

Gallacher, John D, and Rolf E Fredheim. 'Division Abroad, Cohesion at Home: How the Russian Troll Factory Works to Divide Societies Overseas but Spread pro-Regime Messages at Home', in *Responding to Cognitive Security Challenges* (Riga, Latvia: NATO StratCom CoE, 2019), 60–79.

Gallacher, John David, Marc W Heerdink, and Miles Hewstone. "Online Contact between Opposing Political Protest Groups via Social Media Predicts Physical Violence of Offline Encounters." *Under Review*, 1–44.

Gallagher, Ryan J., Andrew J. Reagan, Christopher M. Danforth, and Peter Sheridan Dodds. 'Divergent Discourse between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter', *PLoS ONE* 13, № 4 (2018): 1–23.

Garimella, Kiran, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. 'Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship', *International World Wide Web Conference* 2 (2018).

Garrett, R. Kelly, Brian E. Weeks, and Rachel L. Neo. 'Driving a Wedge Between Evidence and Beliefs: How Online Ideological News Exposure Promotes Political Misperceptions', *Journal of Computer-Mediated Communication* Vol. 21, Issue 5 (2016): 331–48.

Gerber, Theodore P., and Jane Zavisca. 'Does Russian Propaganda Work?', *The Washington Quarterly* Vol. 39, Issue 2 (2016).

Gideon, Lucian, Conway Iii, Kathrene R Conway, and Shannon C Houck. 'Automated Integrative Complexity', *Political Psychology* Vol. 35, № 5 (2014).

Gleicher, Nathaniel. 'Removing Coordinated Inauthentic Behavior from the UK and Romania', Facebook Newsroom, 7 March 2019.

Google Project Jigsaw, 'Perspective', accessed 23 March 2018.

Guess, Andrew, Pablo Barber, Cristian Vaccari, United Kingdom, Brendan Nyhan, Alexandra Seigel, Sergey Sanovich, and Denis Stukal. 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature', The William and Flora Hewlett Foundation, 2018.

Guttieri, Karen, Michael D. Wallace, and Peter Suedfeld. The Integrative Complexity of American Decision Makers in The Cuban Missile Crisis', *Journal of Conflict Resolution* Vol. 39, № 4 (1 December 1995).

Gvirsman, Shira Dvir, R. Kelly Garrett, Aysenur Dal, Rachel Neo, Yariv Tsfati, and Benjamin K. Johnson. 'Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization', *Human Communication Research* Vol. 40, Issue 3 (2014).

Hindman, Matthew, and Vlad Barash. 'Disinformation, "Fake News" and Influence Campaigns on Twitter', 2018.

Hirschfeld, H. O., and J. Wishart. 'A Connection between Correlation and Contingency', *Mathematical Proceedings of the Cambridge Philosophical Society* Vol. 31, Issue 4 (October 1935): 520.

Houck, Shannon C., Meredith A. Repke, and Lucian Gideon Conway. Understanding What Makes Terrorist Groups' Propaganda Effective: An Integrative Complexity Analysis of ISIL and Al Qaeda', *Journal of Policing, Intelligence and Counter Terrorism* Vol. 12, Issue 2 (2017): 105–18.

Houck, Shannon C. 2014. 'Automated Integrative Complexity : Current Challenges and Future Directions', *Political Psychology* 35, Issue 5 (2014): 603–24.

Howard, Philip N, Bharath Ganesh, Dimitra Liotsiu, John Kelly, and Graphika Camille François. 'The IRA, Social Media and Political Polarization in the United States, 2012-2018', University of Oxford Computational Research Project, 2018.

Howell, Lee, Martina N Gmur, Peter Bisanz, Isabel de Sola, Karine Von, Benjamin Prampart, Rigas Hadzilacos, Stefan Hall, and Aude Lanois. 'Outlook on the Global Agenda 2014' (Geneva: World Econoomic Forum, 2014).

Husson, Francois, Julie Josse, Sebastien Le, and Jeremy Mazet. 'Package "FactoMineR"', *CRAN*, 2018.

Intelligence Community Assesment. 'Assessing Russian Activities and Intentions in Recent US Elections', Office of the Director of National Intelligence, 2017.

Karan, Kanisk, Donara Barojan, Melissa Hall, and Graham Brookie. "#TrollTracker: Outward Influence Operation from Iran." 2019.

Kello, Lucas. *The Virtual Weapon and International Order* (Yale University Press, 2017).

King, Gary, Pan Jennifer, and Roberts Margaret E. 'How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument', *American Political Science Review* Vol. 111, Issue 03 (2017): 484–501.

Kirchick, James. 'Russia's Plot against the West', *Politico*, 17 March 2017.

Kosloff, Spee, Jeff Greenberg, Toni Schmader, Mark Dechesne, and David Weise. 'Smearing the Opposition: Implicit and Explicit Stigmatization of the 2008 U.S. Presidential Candidates and the Current U.S. President', *Journal of Experimental Psychology: General* Vol. 139, № 3 (2010): 383–98.

Kumar, Srijan, Justin Cheng, Jure Leskovec, and V. S. Subrahmanian. 'An Army of Me: Sockpuppets in Online Discussion Communities', *Proceedings of the 26th International Conference on World Wide Web* (2017), 857–66.

Lazer, David M. J., Michael Schudson, Yochai Benkler, Jonathan L. Zittrain, Emily A. Thorson, Duncan J. Watts, Matthew A. Baum, et al. 'The Science of Fake News', *Science* Vol. 359, Issue 6380 (2018): 1094–96.

Lee, Eun Ju. 'Deindividuation Effects on Group Polarization in Computer-Mediated Communication: The Role of Group Identification, Public-Self-Awareness, and Perceived Argument Quality', *Journal of Communication* Vol. 57, Issue 2 (2007): 385–403.

Lim, Ee-Peng, Palakorn Achananuparp, and Feida Zhu. 'On Modeling Virality of Twitter Content', *ICADL*, 2011, 212:221.

Linvill, Darren L, and Patrick L Warren. Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building', *(In press)*.

Lucas, Edward, and Peter Pomerantsev. 'Winning the Information War: Techniques and Counter-Strategies to Russian Propaganda in Central and Eastern Europe', Center for European Policy Analysis & The Legatum Institute, 2016.

Maechler, Martin. 'Package "Diptest"', *CRAN*, 5 December 2015.

Margolin, Drew B., Aniko Hannak, and Ingmar Weber. 'Political Fact-Checking on Twitter: When Do Corrections Have an Effect?', *Political Communication* Vol. 35, 2 (2018): 196–219.

Mason, Lilliana. '"I Disrespectfully Agree": The Differential Effects of Partisan Sorting on Social and Issue Polarization', *American Journal OfPolitical Science* Vol. 59, Issue 1 (2014): 128–45.

Metaxas, Panagiotis, and Twittertrails Research Team. 'Retweets Indicate Agreement, Endorsement, Trust: A Meta-Analysis of Published Twitter Research', *ArXiv Preprint*, 2017.

Mihaylov, Todor, Georgi Georgiev, and Preslav Nakov. 'Finding Opinion Manipulation Trolls in News Community Forums', *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, № July (2015): 310–14.

Müller, Karsten, and Carlo Schwarz. 'Fanning the Flames of Hate: Social Media and Hate Crime', *SSRN Electronic Journal*, 7 December 2017.

Muralidharan, Sidharth, and Linjuan Rita Men. 'How Peer Communication and Engagement Motivations Influence Social Media Shopping Behavior: Evidence from China and the United States', *Cyberpsychology, Behavior, and Social Networking* Vol. 18, № 10 (2015): 595–601.

Newman, Nic, and David a. L. Levy. 'Reuters Institute Digital News Report 2017' (Reuters Institute for the Study of Journalism, 2017).

Nickerson, Raymond S. 'Confirmation Bias: A Ubiquitous Phenomenon in Many Guises', *Review of General Psychology* Vol. 2, № 2 (1998): 175–220.

Pariser, Eli. *The Filtter Bubble: What the Internet Is Hiding from You* (New York: Penguin Press, 2011).

Paul, Christopher, and Miriam Matthews. "'The Russian 'Firehose of Falsehood' Propaganda Model" 2016.

Pennycook, Gordon, Tyrone D Cannon, and David G Rand. 'Prior Exposure Increases Perceived Accuracy of Fake News', *Journal of Experimental Psychology* Vol. 147, № 12 (2018): 1865–80.

Pettigrew, Thomas F, and Linda R Tropp. 'How Does Intergroup Contact Reduce Prejudice? Meta-Analytic Tests of Three Mediators', *European Journal of Social Psychology* Vol. 38, Issue 6 (2008): 922–34.

Postmes, Tom, Russell Spears, and Martin Lea. 'Building or Breaching Social Boundries? SIDE Effects of Computer Mediated Communication', *Communication Research* 25, № 6 (1998).

Preoțiuc-Pietro, Daniel, Ye Liu, Daniel Hopkins, and Lyle Ungar. 'Beyond Binary Labels: Political Ideology Prediction of Twitter Users', *Proceedings Ofthe 55th Annual Meeting of the Association for Computational Linguistics*, 2017, 729–40.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 'Language Models Are Unsupervised Multitask Learners', 2018.

Rani, Neeta. 'Social Media in India: A Human Security Perspective', *The Research Journal of Social Sciences* Vol. 9, № 10 (2018): 43–52.

Reddit. Com Traffic, Demographics and Competitors—Alexa', accessed 25 February 2019.

Reddit. Reddit's 2017 Transparency Report and Suspect Account Findings', accessed 25 February 2019.

Roozenbeek, Jon, and Sander Van Der Linden. 'The Fake News Game: Actively Inoculating Against the Risk of Misinformation', *Journal of Risk Research* 9877 (2018): 1–11.

Salihefendic, Amir. 'How Reddit Ranking Algorithms Work', Hacking and Gonzo, *Medium*, 8 December 2015.

Sanovich, Sergey. 'Computational Propaganda in Russia: The Origins of Digital Misinformation', University of Oxford Computational Propaganda Research Project, 2017.

Shearer, Elisa, and Jeffrey Gottfried. 'News Use across Social Media Platforms 2017', Pew Research Center, 17 September 2017.

Shin, Jieun, Lian Jian, Kevin Driscoll, and François Bar. 'Political Rumoring on Twitter during the 2012 US Presidential Election: Rumor Diffusion and Correction', *New Media and Society* Vol. 19, № 8 (2017).

Singer, P. W. (Peter Warren), and Emerson T. Brooking. *Likewar: The Weaponization of Social Media* (New York: Houghton Mifflin Harcourt Publishing Company, 2018).

Smith, Allison, Peter Suedfeld, Lucian Conway, and David Winter. 'The Language of Violence: Distinguishing Terrorist from Nonterrorist Groups by Thematic Content Analysis', *Dynamics of Asymmetric Conflict* Vol. 1, Issue 2 (July 2008): 142–63.

Sobolev, Anton. 'Fantastic Beasts and Whether They Matter: How pro-Government "Trolls" Influence Political Conversations in Russia', (*In Prep*).

Statista. 'Number of monthly active Twitter users worldwide from 1st quarter 2010 to 1st quarter 2019 (in millions)', 2019, accessed February 25, 2019.

Stecklow, Steve. 'Why Facebook Is Losing the War on Hate Speech in Myanmar', *Reuters*, 15 August 2018.

Stewart, Leo G., Ahmer Arif, and Kate Starbird. 'Examining Trolls and Polarization with a Retweet Network', *Proceedings of WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2).*

Stewart, Leo Graiden, Ahmer Arif, A. Conrad Nied, Emma S. Spiro, and Kate Starbird. 'Drawing the Lines of Contention: Networked Frame Contests within #BlackLivesMatter Discourse', *Proceedings of the ACM on Human-Computer Interaction* Vol. 1, Issue CSCW, Article № 96 (2017): 1–23.

Streufert, S, and P Suedfeld. 'Conceptual Structure, Information Search, and Information Utilization', *Journal of Personality and Social Psychology* Vol. 2, № 5 (November 1965): 736–40.

Suedfeld, Peter, and Susan Bluck. 'Changes in Integrative Complexity Prior to Surprise Attacks', *Journal of Conflict Resolution* Vol. 32, № 4 (1988): 626–35.

Sunstein, Cass R. *#Republic: Divided Democracy in the Age of Social Media* (Princeton University Press, 2017).

Tetlock, Philip E., Randall S. Peterson, and Jane M. Berry. 'Flattering and Unflattering Personality Portraits of Integratively Simple and Complex Managers', *Journal of Personality and Social Psychology* Vol. 64, № 3 (1993): 500–11.

Turner, J. C., B. Davidson, and M. A. Hogg. 'Polarized Norms and Social Frames of Reference: A Test of the Self-Categorization Theory of Group Polarization', *Basic and Applied Social Psychology* Vol. 11, № 1 (1990): 77–100.

UK Department for Digital Culture Media and Sport Committee. 'Disinformation and "Fake News": Interim Report', 2018.

UK Department for Digital Culture Media and Sport Committee. Disinformation and "Fake News": Final Report', 2019.

Vicario, Michela Del, Sabrina Gaito, Walter Quattrociocchi, Matteo Zignani, and Fabiana Zollo. 'Public Discourse and News Consumption on Online Social Media: A Quantitative, Cross-Platform Analysis of the Italian Referendum', arXiv.org, February 2017.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 'The Spread of True and False News Online', *Science* Vol. 359, Issue 6380 (2018).

Weedon, Jen, William Nuland, and Alex Stamos. '*Information Operations and Facebook'* Facebook, 2017.

DOI 10.30966/2018.RIGA.6.5.

Weeks, Brian E. 'Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation', *Journal of Communication* vol. 65, Issue 4 (2015): 699–719.

Woolley, Samuel C, and Philip N Howard. 'Political Communication, Computational Propaganda, and Autonomous Agents' *International Journal of Communication*, 10 (2016): 4882–90.

Wright, Jordan, and Olabode Anise. 'Don't @ Me : Hunting Twitter Bots at Scale', *Black Hat*, 2018, 1–43.

Wulczyn, Ellery, Nithum Thain, and Lucas Dixon. 'Ex Machina: Personal Attacks Seen at Scale', *Proceedings of the 26th International Conference on World Wide Web*, 2017, 1391–99.

Yardi, Sarita, and Danah Boyd. 'Dynamic Debates: An Analysis of Group Polarization over Time on Twitter', *Bulletin of Science, Technology & Society* 30, № 5 (2010): 316–27.

Zannettou, Savvas, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 'Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter', (2019).

Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 'Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web', (2018).

Zannettou, Savvas, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 'Who Let the Trolls out? Towards Understanding State-Sponsored Trolls', (2019)

Zhou, Lu, Wenbo Wang, and Keke Chen. 'Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones', *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 2016, 603–12.