"Donald Trump is my President!"
The Internet Research Agency Propaganda Machine

Marco Bastos (@toledobastos) - *City, University of London*
Johan Farkas (@farkasjohan) - *Malmö University*

Accepted for publication in Social Media + Society
(pre-publication version: some changes still possible)

Abstract

This article presents a typological study of the Twitter accounts operated by the Internet Research Agency (IRA), a company specialized in online influence operations based in St. Petersburg, Russia. Drawing on concepts from 20th century propaganda theory, we modeled the IRA operations along propaganda classes and campaign targets. The study relies on two historical databases and data from the Internet Archive's Wayback Machine to retrieve 826 user profiles and 6377 tweets posted by the agency between 2012 and 2017. We manually coded the source as identifiable, obfuscated, or impersonated and classified the campaign target of IRA operations using an inductive typology based on profile descriptions, images, location, language, and tweeted content. The qualitative variables were analyzed as relative frequencies to test the extent to which the IRA's black, grey, and white propaganda are deployed with clearly-defined targets for short, medium, and long-term propaganda strategies. The results show that source classification from propaganda theory remains a valid framework to understand IRA's propaganda machine and that the agency operates a composite of different user accounts tailored to perform specific tasks, including openly pro-Russian profiles, local American and German news sources, pro-Trump conservatives, and black lives matter activists.

Keywords: Propaganda; Internet Research Agency; Trolls; Disinformation; Twitter

Introduction

Disguised propaganda was a stable part of military operations in the 20th century as means of weakening enemy states (Linebarger, 1954). The aftermath of the Cold War was nonetheless marked by a declining trend in information warfare between enemy states, a drift shadowed by the waning importance of propaganda studies in the period (Briant, 2015). This uncontroversial assessment was recently challenged in the aftermath of the US Presidential election of 2016 and the United Kingdom's referendum on EU membership, with multiple reports of social media platforms being weaponized to spread hyperpartisan content and propaganda (Bastos & Mercea, 2019; Bessi & Ferrara, 2016). This study seeks to further explore the weaponization of social media platforms by inspecting 826 Twitter accounts and 6377 tweets created by the Kremlin-linked Internet Research Agency (IRA) in St. Petersburg.

Propaganda studies classify manipulation techniques according to different source classes. White propaganda refers to unambiguous, openly identifiable sources in sharp contrast to black propaganda in which the source is disguised. Grey propaganda sits somewhere in between these classes with the source not being directly credited nor identified (Becker, 1949; Doherty, 1994; McAndrew, 2017). Propaganda models are however reminiscent from a media ecosystem dominated by mass media and broadcasting. As such, the classic propaganda models probe into processes of framing, priming, and schemata, along with a range of media effects underpinning information diffusion in the postwar period leading up to the Cold War (Hollander, 1972), but invariable predating the internet (Hermans, Klerkx, & Roep, 2015).

We probe the propaganda efforts led by the Internet Research Agency, a so-called "troll factory" reportedly linked to the Russian government (Bertrand, 2017), by relying on a list of deleted Twitter accounts that was handed over to the U.S. Congress by Twitter on 31 October 2017 as part of their investigation into Russia's meddling in the 2016 U.S. elections (Fiegerman & Byers, 2017). According to Twitter, a total of 36,746 Russian accounts produced approximately 1.4 million tweets in connection to the U.S. elections (Bertrand, 2017). Out of these accounts, Twitter established that 2752 were operated by the IRA (United States Senate Committee, 2017). In January 2018, this list was expanded to include 3814 IRA-linked accounts (Twitter, 2018b).

The messages explored in this study were posted between 2012 and 2017 by IRA-linked accounts. We employ a mixed-methods approach to retrieve, analyze, and manually code 826

Twitter accounts and 6377 tweets from the IRA that offer insights into the tactics employed by foreign agents engaging in "information warfare against the United States of America" (US District Court, 2018, p. 6). Drawing on source classification from propaganda studies, we detail IRA's tactical operationalization of Twitter for disguised propaganda purposes. In the following, we review the literature on propaganda studies and present an overview of what is currently known about the IRA's disinformation campaigns. We subsequently explore the differences between white, grey, and black propaganda distributed by the IRA with clearly-defined campaign targets. We expect the relationship between campaign target and propaganda classes to reveal IRA's operational strategies and campaign targets.

Previous work

Propaganda and information warfare have traditionally been studied in the context of foreign policy strategies of nation states, with mass media such as newspapers, radio, and television sitting at the center of disinformation campaigns (Jowett & O'Donnell, 2014). In fact, mass media and propaganda techniques evolved together in the 20th century towards a state of global warfare (Cunningham, 2002; Taylor, 2003). During this period, both the definition and forms of propaganda changed dramatically (Welch, 2013), but the centrality of mass media remained a relatively stable component of propaganda diffusion (Cunningham, 2002), a development captured by Ellul (1965) who argued that modern propaganda could not exist without the mass media. Towards the end of the 20th century, where media plurality increased dramatically through the rise of cable TV and the Internet, propaganda operations were seen as a remnant of the past and largely abandoned in scholarly literature (Cunningham, 2002). Combined with the end of the Cold War, propaganda was broadly seen as both technologically and politically outdated.

The notion that increased media diversity made large-scale propaganda campaigns obsolete continued with the rise of social platforms, enabling citizens and collectives to produce counter-discourses to established norms, practices, and policies. Boler and Nemorin (2013, p. 411) reflected this optimism by arguing that "the proliferating use of social media and communication technologies for purposes of dissent from official government and/or corporate-interest propaganda offers genuine cause for hope." By the end of the decade, however, this sentiment had changed considerably as the decentralized structure of social media platforms

enabled not only public deliberation, but also the dissemination of propaganda. Large-scale actors such as authoritarian states sought to coordinate propaganda campaigns that appeared to derive from within a target population, often unaware of the manipulation (US District Court, 2018). The emergence of social network sites thus challenged the monopoly enjoyed by the mass media (Castells, 2012), but it also offered propagandists a wealth of opportunities to coordinate and organize disinformation campaigns through decentralized and distributed networks (Benkler, Faris, & Roberts, 2018).

Upon the consolidation and the ensuing centralization of social platforms, state actors efficiently appropriated social media as channels for propaganda, with authoritarian states seizing the opportunity to enforce mass censorship and surveillance (Khamis, Gold, & Vaughn, 2013; King, Pan, & Roberts, 2017; Youmans & York, 2012). Technological advances in software development and machine learning enabled automated detection of political dissidents, removal of political criticism, and mass dissemination of government propaganda through social media. These emerging forms of political manipulation and control constitute a difficult object of analysis due to scant and often non-existing data, largely held by social media corporations that hesitate to provide external oversight to their data (Bastos & Mercea, 2018b) while offering extensive anonymity for content producers and poorly handling abusive content (Farkas, Schou, & Neumayer, 2018).

In the context of the 2016 UK EU membership referendum, research estimates that 13,493 Twitter accounts comprised automatic posting protocols or social bots—i.e., software-driven digital agents producing and distributing social media messages (Bastos & Mercea, 2019). By liking, disseminating, and re-tweeting content, these accounts collectively produced 63,797 tweets during the referendum debate. In the US context, Bessi and Ferrara (2016) used similar bot-detection techniques to find 7183 Twitter accounts that tweeted the 2016 US elections and similarly displayed bot-like characteristics. Despite the reported high incidence of bot activity on social media platforms, researchers can only identify bot-like accounts retrospectively based on their activity patterns and characteristics that set them apart from human-driven accounts, most prominently the ratio of tweets to retweets, which is higher for social bots (Bastos & Mercea, 2019).

Establishing the identity of content producers in the social supply chain is challenging in cases of disguised social media accounts. While social bots can be identified based on traces of

computer automation, disguised human-driven accounts can be difficult to recognize because they lack unambiguous indicators of automation. Disguised human-driven accounts can neither be easily found nor traced back to an original source or controller. Reliable identification of such accounts requires collaboration with social media companies which are reluctant to provide such support (Hern, 2017). In fact, the list of 3814 deleted accounts identified as linked to the IRA and explored in this study was only made public by Twitter by request of the US Congress (United States Senate Committee, 2017).

Disguised propaganda and Information Warfare

Jowett and O'Donnell (2014, p. 7) define propaganda as the "deliberate, systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist." Propaganda campaigns are often implemented by state actors with the expectation of causing or enhancing information warfare (Jowett & O'Donnell, 2014; Linebarger, 1948). Unlike propaganda targeted at a state's own population, information warfare is waged against foreign states and it is not restricted to periods of armed warfare; instead, these efforts "commence long before hostilities break out or war is declared… [and] continues long after peace treaties have been signed" (Jowett & O'Donnell, 2014, p. 212). After the end of the Cold War, the concepts of propaganda and information warfare were perceived as anachronistic and rapidly abandoned in scholarly discourse (Winseck, 2008, p. 421). With the recent rise of large-scale information campaigns and infiltration through digital media platforms, scholars are nonetheless increasingly arguing for the continued relevance of propaganda theory (Benkler, et al., 2018; Farkas & Neumayer, 2018; Woolley & Howard, 2019). Western democratic and military organizations likewise restored the notion of "information warfare" in the context of military build-ups between Russia and NATO allies, particularly the US (Giles, 2016; US District Court, 2018).

A key objective of information warfare is to create confusion, disorder, and distrust behind enemy lines (Jowett & O'Donnell, 2014; Taylor, 2003). Through the use of grey or black propaganda, conflicting states have disseminated rumors and conspiracy theories within enemy territories for "morale-sapping, confusing and disorganizing purposes" (Becker, 1949). Within propaganda theory, grey propaganda refers to that which has an unidentifiable or whose source is difficult to identify, while black propaganda refers to that which claims to derive from within the

enemy population (Daniels, 2009; Jowett & O'Donnell, 2014). As noted by Daniels (2009), this source classification model is problematic due to its racial connotations, but the distinction between identifiable (white propaganda), unidentifiable (grey propaganda), and disguised sources (black propaganda) has been effectively used to analyze different types of information warfare throughout the 20th century.

Internet Research Agency and the Kremlin connection

The Internet Research Agency (IRA) is a secretive private company based in St. Petersburg reportedly orchestrating subversive political social media activities in multiple European countries and the US, including the 2016 US elections (Bugorkova, 2015; The Economist, 2018). The US District Court (2018) concluded that the company engages in "information warfare" based on "fictitious U.S. personas on social media platforms and other Internet-based media." The court also linked the IRA to the Russian government through its parent company, which holds various contracts with the Russian government. There is also evidence linking the founder of IRA, Yevgeniy Prigozhin, to the Russian political elite. The Russian government has nonetheless rejected accusations of involvement in subversive social media activities and downplayed the US indictment of Russian individuals (MacFarquhar, 2018).

The IRA has been dubbed a "troll factory" due to its engagement in social media trolling and the incitement of political discord using fake identities (Bennett & Livingston, 2018). This term has clear shortcomings, as the agency's work extends beyond trolling and includes large-scale subversive operations. According to internal documents leaked in the aftermath of the US election, the workload of IRA employees was rigorous and demanding. Employees worked 12-hour shifts and were expected to manage at least six Facebook fake profiles and ten Twitter fake accounts. These accounts produced a minimum of three Facebook posts and 50 tweets a day (Seddon, 2014). Additional reports on the subversive operations of the IRA described employees writing hundreds of Facebook comments a day and maintaining several blogs (Bugorkova, 2015). These activities were aimed at sowing discord among the public. In the following, we unpack our research questions and our methodological approach, including the challenges posed by data collection and retrieval and the study of obfuscated and impersonated Twitter accounts.

Research questions and hypotheses

This study is informed by propaganda studies and examines a number of exploratory hypotheses regarding the tactics and use of disguised propaganda on Twitter. Our first hypothesis draws from Becker (1949, p. 1) who argued that black propaganda is an effective means of information warfare in contexts of "widespread distrust of ordinary news sources." This is in line with reports of falling trust in the press, with only 33% of Americans, 50% of Britons, and 52% of Germans trusting news sources (Newman, Richard Fletcher, Levy, & Nielsen, 2016). To this end, we hypothesize that IRA-linked Twitter accounts will leverage the historical low level of trust in the media and deploy mostly black propaganda (H1a) as opposed to grey (H1b) or white (H2c) propaganda.

Secondly, we hypothesize that Russian propaganda is aimed at spreading falsehoods and conspiracy theories to drive a wedge between groups in the target country. This is consistent with traditional propaganda classes, so hypothesis H2 tests whether black propaganda fosters confusion and stokes divisions by spreading fearmongering stories, relies on expletives and hostile expression, and disseminates populism appeals that position "the people" against the government (H2a); or, alternatively, whether this type of content is disseminated by employing (H2b) grey or (H2c) white propaganda (Jowett & O'Donnell, 2014).

Thirdly, we explore the mechanisms through which the IRA has engaged in subversive information warfare, which often comes in the form of propaganda of agitation disseminated to stir up tension through the use of "the most simple and violent sentiments… hate is generally its most profitable resource" (Ellul, 1965). Following this seminal definition provided by Ellul (1965), we seek to test whether IRA propaganda on social media promotes agitation, emotional responses, direct behavior, polarization, and support for rumors and conspiracy theories by strategically deploying black (H3a), grey (H3b), or white (H3c) propaganda to disseminate these sentiments, expressions, and stories.

Fourth, we rely on an inductive typology of Twitter accounts to explore the IRA propaganda strategy across a range of targets, including protest activism (e.g., Black Lives Matter), local news diffusion, and conservative ideology. To this end, we convert the typology to a numeric variable and test whether the strategic target of IRA campaigns is associated with and predictive of propaganda type (H4). Lastly, we unpack this relationship by exploring the temporal patterns associated with propaganda classes and campaign targets.

Methods and Data

Data Collection

Investigating the cohort of 3814 IRA accounts was challenging, as Twitter did not share deleted tweets and user profiles with researchers and journalists until October 2018—two years after the US elections and a full year after the company admitted to Russian interference (Gadde & Roth, 2018). In addition to that, Twitter policy determines that content tweeted by users should be removed from the platform once the account is deleted or suspended (Twitter, 2018a). As a result, the tweets posted by the 3814 IRA accounts are no longer available on Twitter's Search, REST, or Enterprise APIs.

To circumvent this limitation, we first queried a large topic-specific historical Twitter database spanning 2008-2017. This database spanned a range of topics from our previous studies on U.S. daily news consumption dating back to 2012 (Bastos & Zago, 2013), Brazilian and Ukrainian protests in 2013 and early 2014 (Bastos & Mercea, 2016), the Charlie Hebdo terrorist attack in 2015, and the Brexit referendum in 2016 (Bastos & Mercea, 2018a). We found evidence of IRA interference across most of the data. In this first step of data collection, we retrieved 4989 tweets posted by IRA accounts from the historical datasets. NBC News subsequently published a dataset of over 200,000 tweets from 454 IRA accounts curated by anonymous researchers (Popken, 2018). The distribution of messages in this dataset is fairly skewed, with 140 users having tweeted less than 10 messages and 27 accounts having tweeted over 3000 messages. We nonetheless sampled 10 tweets from each account in this database (if available), thus retrieving 1388 and expanding our coded dataset to 6377 tweets.

Lastly, we queried the Wayback Machine API and found 102 user profiles available in the Internet Archive. Only a few snapshots included tweeted content, so we relied on Wayback Machine as a source of user profile, which is the unit of analysis in this study. The aggregate database explored in this study thus consists of 826 user profiles and 6377 tweets posted by IRA-linked accounts, which translates to just over one-fifth of the accounts identified by Twitter as linked to the IRA (21.7% of 3814). The database comprises 15 variables for each account, including the textual variables username, user ID, self-reported location, account description, and website; the numeric variables account creation date, number of tweets and favorited tweets by the account, and the number of followers and followees; and logical or binary variables

indicating whether the account is verified and protected. The database is text-only and therefore we do not have access to images or videos embedded to the tweets created by IRA sources.

Coding and analysis

Tweets were manually and systematically annotated by an expert coder along 18 variables, 17 of which were established deductively. The 18th variable identifies the most prominent issues mentioned by the account and was established inductively based on an initial coding of a subsample of 10% of tweets. A total of 15 issues were identified as deductive attributes upon coding the dataset. In order to ensure consistency, a codebook describing each variable and attribute was used throughout the coding (see appendix). Variables are not mutually exclusive, nor do they apply to all tweets in the dataset. The manual coding took around 175 hours and an overview of the variables for tweets and accounts is presented in Table 1.

Each IRA account was coded based on three variables: user type, national identity, and campaign target. Campaign target was established by training a set of 250 accounts (30% of accounts) to render a typology of campaign targets of IRA-operated accounts in our database. The typology was created based on recurrent identifiers in account descriptions, language, time zone, nationality, and tweeted content. Five broad campaign targets were identified, each containing a number of sub-targets: *Russian citizens* (including Russian politics, Russian news, and self-declared Russian propagandists); *Brexit* (including mainstream media coverage and support to the Brexit campaign); *Conservative patriots* (including Republican content); *Protest activism* (including Black Lives Matter, Anti-Trump, and Anti-Hillary communication); and *Local news*, whose accounts mostly post and retweet mainstream media sources.

We relied on the typologies described above to generate dependent and independent variables guiding this study. The dependent variables are propaganda classes and campaign targets. Propaganda classes are divided in identifiable, obfuscated, and impersonated. Campaign targets comprise conservative patriots, local news, protest activists, and Brexit. The independent variables were calculated by normalizing and subsequently quantifying the instances of fearmongering, populist sentiment, emotional charge, polarization, hostility, and conspiracy-theorizing associated with each IRA-linked account. These variables are analyzed in reference to user accounts, which is the unit of analysis underpinning our study.

Table 1: Manually coded variables for IRA tweets and accounts (see appendix for additional information)

Coding variables: tweets

1. National context of tweet drawn from content
2. Language
3. Retweeted Twitter account
4. Mentioned or replied Twitter account
5. Mentioned person or organization (non-Twitter user)
6. Political party mentioned, retweeted, or replied to (person or account).
7. Endorsement of individual, organization, or cause
8. Disapproval of individual, organization, or cause
9. Religion
10. Fatalities ('risk of fatality', 'fatality', 'fatalities', '5+ fatalities' and 'mass murder')
11. Rumor/conspiracy theory ('yes' and 'high')
12. Aggressiveness ('yes' and 'high').
13. Antagonism ('yes' and 'high').
14. Emotional ('yes' and 'high').
15. Encouragement of action ('vote X' or 'share this!')
16. Populist rhetoric (reference to 'the people', 'anti-establishment', 'anti-mainstream media', 'scapegoating', 'call for action', 'ethno-cultural antagonism', 'state of crisis/threat against society', 'the need for a strong leader')
17. Populism spectrum (two attributes: 'low' and 'high')'
18. Issues (up to four attributes per tweet based on 15 attributes established through an inductive coding of a sub-set of 10% of tweets)

Coding variables: accounts

19. User type (eight attributes, including 'individual (male)', 'individual (female)', 'news source' and 'NGO')
20. National user identity (based on declared location, time zone, and self-description in user profile and tweets)
21. Campaign target (based on five overall attributes established through an inductive coding of a sub-set of 30% of accounts)

In summary, the qualitative variables assigned to tweets were subsequently converted to numeric and logical scales for hypothesis testing. The variable fearmongeringScore was created by calculating the average number of tweets and news articles mentioning fatalities caused by natural disasters, crime, acts of terrorism, civil unrest, or accidents. We assign a value of zero to tweets with no such mention, 1 when the risk of fatality is mentioned, 2 for direct mentions of fatality, 3 for multiple fatalities, 4 for reports of five or more fatalities, and 5 for mass murders and military conflicts with several casualties. The variable populistScore was calculated by assigning a scale of 0 to 3 based on the incidence of messages appealing, among other things, to "the people" in their struggle against a privileged elite (Mudde, 2004). Emotional messages were coded in a scale from 0 for not emotional to 2, with messages scoring 2 having the highest levels of emotional content. A similar scale was applied to variables "antagonism" and "aggressiveness," with 0 for no such sentiment, 1 for positive matches, and 2 for messages with high incidence of said content. We follow similar scales for variables rumor and conspiracy theory (6 scales) and the encouragement of offline action. This procedure enabled us to identify the propaganda class of each account and six numeric variables that measure the levels of fearmongering, populist sentiment, emotional charge, polarization, hostility, conspiracy-theorization, and incitement to offline action associated with that account.

Limitations of the Method and Data

The disguised propaganda produced by the IRA and explored in this study has been retrieved by trawling through millions of previously archived tweets to identify messages authored by the 3814 accounts Twitter acknowledged as operated by the IRA. One account turned out to be a false-positive and was excluded from the study (Matsakis, 2017). The dataset spans eight years and includes tweets with a topical focus on US news outlets, the Charlie Hebdo terrorist attack in 2015, and the Brexit debate in 2016.

A portion of the database was encoded in Latin-1 Supplement of the Unicode block standard, which does support Cyrillic characters, hence messages in Russian or Ukrainian could not be annotated. A total of 1848 tweets posted during the Euromaidan wave of demonstrations and civil unrest in Ukraine were encoded in the Cyrillic alphabet and the tweets could not be annotated because they did not include text. We relied on the profile retrieved for these users to classify them as Russian, and thus as white propaganda, as Twitter already identified them as IRA-linked. We nonetheless acknowledge that the absence of tweets for this cohort of accounts impinge on our ability to identify them as sources impersonating Ukrainian as opposed to Russian users, in which case the incidence of black propaganda would be considerably higher than identified in this study.

The data explored in this study represents only a portion of the IRA propaganda efforts. Accordingly, our study cannot estimate the extent of IRA propaganda on social media nor the prevalence of other forms of propaganda tactics. Similarly, the inductive typology employed in this study does not necessarily comprehend the totality of strategies deployed by the IRA. Lastly, and contrary to our expectations, we identified several pro-Russia accounts claiming to be "run by the Kremlin." While it is not possible to determine the extent to which the Russian government was involved in the IRA operations, for the purposes of this study we consider these accounts as Russian and therefore as sources of white propaganda.

Results

The summary statistics allow us to approach hypothesis H1 by inspecting the breakdown of IRA-linked Twitter accounts dedicated to black, grey, and white propaganda. We find that most accounts operated by the Internet Research Agency are dedicated to disseminating black propaganda (42%, n= 339), followed by white (40%, n= 319) and grey (18%, n=141)

propaganda. Similarly, the sample of manually coded tweets follows a comparable distribution, with 58% (*n*=3450) of messages coded as black propaganda as opposed to grey (5%, *n*=321) or white propaganda (37%, *n*=2205). The distribution of tweets, followers, and followees lend further support to H1a, as black propaganda accounts present more capillarity with higher number of followers, followees, and average number of messages posted by these accounts compared with grey and white accounts. Figure 1 unpacks the differences across classes.

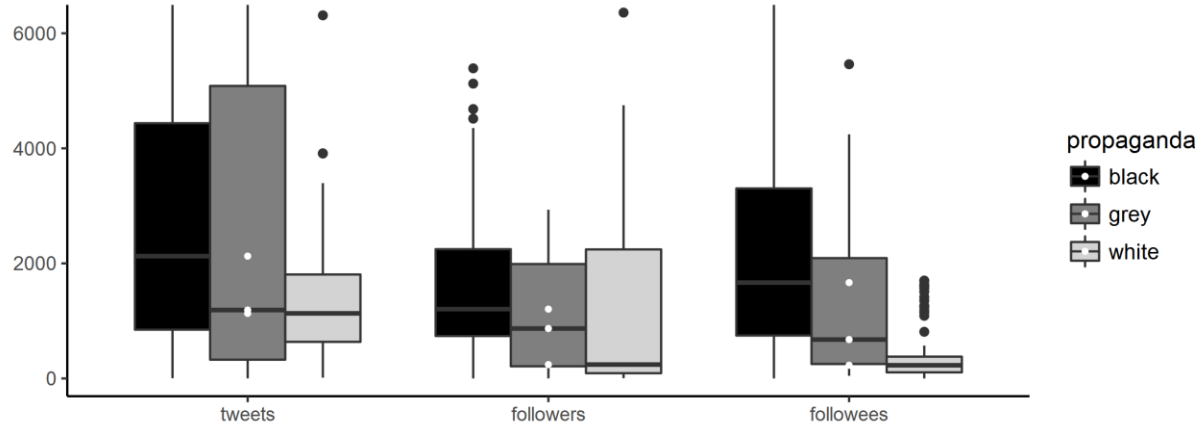

Figure 1: Number of tweets, followers, and followees for accounts dedicated to black, grey, and white propaganda

We subsequently test hypothesis H2, which hypothesized that IRA efforts to spread falsehoods and conspiracy theories would be segmented across propaganda classes, tailored to wedge divisions in the target country. The data lend support for hypothesis H2(b), with grey propaganda scoring consistently higher than black and white for fearmongering ($\bar{x}$=.55, .10, .04, respectively), populism sentiments ($\bar{x}$=.35, .19, .02, respectively), and hostility ($\bar{x}$=.22, .15, .01, respectively). The results thus confound our expectations, as the Internet Research Agency seems to favor accounts with unidentifiable location and whose affiliation is concealed to disseminate fearmongering, populist appeals, and hostile political platforms, including scapegoating and call for action against threats to society.

Hypothesis H3 was approached by probing IRA-linked profiles dedicated to emotion-charged stories, polarized political commentary, and the spreading of rumors and conspiracy stories. We assign a score to each category and calculate the mean and standard deviation across propaganda classes. The data lends support to hypothesis H3a, as black propaganda accounts show consistently higher scores for each of the variables tested, particularly emotionScore and

polarizedScore, which averaged .53 and .37 for black propaganda compared with 0.39 and .30 for grey, and .08 and 0.5 for white propaganda. This pattern also holds for the variable measuring posting behavior supporting conspiracy theories, which averaged .46 for black propaganda compared with .29 and .09 for grey and white propaganda accounts, respectively. Figure 2 shows the breakdown across classes.

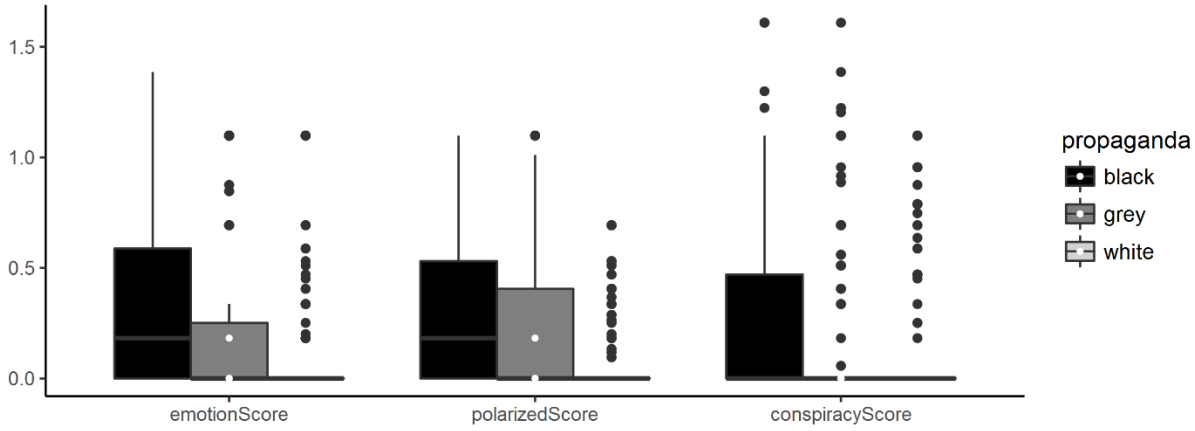

Figure 2: Breakdown of emotional response, polarization, and conspiracy theorizing across propaganda classes

The results indicate that grey propaganda is preferred to disseminate fearmongering stories, stoking populism sentiments, and encouraging hostile expression. Black propaganda, on the other hand, is central to efforts of sowing social discord in the target population. These two classes of propaganda were used to stoke fears in the public and they contrast with self-identified Russian accounts that tweet mostly pro-Kremlin content. Indeed, the mean score of fearmongerScore, populistScore, emotionScore, polarizedScore, hostilityIndex, conspiracyScore, and behaviorIndex are significantly higher in black (.10, .19, .53, .37, .15, .46, .12) and grey (.55, .36, .39, .30, .22, .28, and .20) propaganda compared with white propaganda, which displays low levels of such sentiments (.04, .02, .08, .08, .01, .08, and .05). Fifteen percent of black and 36% of grey propaganda accounts engaged in fear mongering compared with only 6% of white propaganda accounts. Similarly, 26% of grey and 20% of black accounts tweeted populist appeals compared with 2% of white accounts. The trend continues for emotionScore (grey=27%, black=54%, white=9%), polarizedScore (grey=34%, black=57%, white=10%), hostilityIndex (grey=15%, black=25%, white=2%), conspiracyScore (grey=23%, black=45%, white=8%), and behaviorIndex (grey=22%, black=21%, white=5%)

We further delve into hypothesis H3 by performing a stepwise model selection by Akaike Information Criterion (AIC) to predict account type (black, grey, or white). The returned stepwise-selected model includes an ANOVA component that rejects a range of numeric variables, including the number of tweets posted by users and the number of lists associated with the account, but that incorporates all variables coded for this study. Therefore, the model includes fearmongerScore, populistScore, emotionScore, polarizedScore, hostilityScore, conspiracyScore, and behaviorScore, with polarizedScore and conspiracyScore being particularly significant predictors of account type. The model accounts for nearly half of the variance in the data (R^2_{adj}=.40, p=6.836e-15). The results lend support to the hypothesis that source classification remains a valid framework to understand IRA's social media operations, as account type is significantly associated with the dissemination of polarizing, populist, fear mongering, and conspiratorial content.

Table 2: Contingency table of campaign targets by propaganda classes. Local news outlets include the accounts impersonating local news (black propaganda) and accounts dedicated to retweeting this content (grey propaganda)

	Black	**Grey**	**White**
Brexit	49	14	3
Conservative patriots	74	1	1
Local news outlets	45	59	0
Protest activism	72	11	0
Russian/Ukrainian issues	1	0	36

Lastly, we approach hypothesis H4 by inductively coding a typology of IRA Twitter accounts based on their target campaigns, including protest activism (e.g., Black Lives Matter), local news diffusion, and conservative ideology. To this end, we convert the typology to a numeric variable and test whether the propaganda classes are associated with and predictive of the IRA campaign targets. As shown in Table 2, propaganda classes appear dedicated to specific campaigns, with grey propaganda dedicated to local news and the Brexit campaign, black propaganda deployed across campaign targets, and white propaganda unsurprisingly covering Russian and potentially Ukrainian issues almost exclusively. We subsequently performed another stepwise model selection including the campaign target variable, which was found to be a strong predictor of propaganda type. Indeed, most variables previously found to be significant were discarded in the

Stepwise Model Path and only the variables populistScore, emotionScore, polarizedScore, conspiracyScore, and campaign target were deemed relevant predictors of account type (R^2_{adj}=.55, p=2.155e-12). The results are thus consistent with hypothesis H4 and show that source classification from propaganda theory is significantly associated with campaign targets.

The temporal patterns associated with creation and deployment of propaganda accounts add further evidence to the strategic deployment of IRA "trolls." White accounts were largely created and deployed in a timeline that mirrors the Euromaidan demonstrations and the civil unrest in Ukraine in late 2013 and the ensuing annexation of the Crimean Peninsula in early 2014. White accounts were often openly pro-Kremlin and tweeted mostly in Russian and Ukrainian, another marker of the geographic and linguistic boundaries of this operation. Indeed, nearly 70% of white propaganda accounts were created between 2013 and 2014 and nearly 80% of the tweets posted by these accounts took place in 2014 following the annexation of Crimea. Figure 3 unpacks the relationship between account creation date and activity patterns for black, grey, and white propaganda accounts.

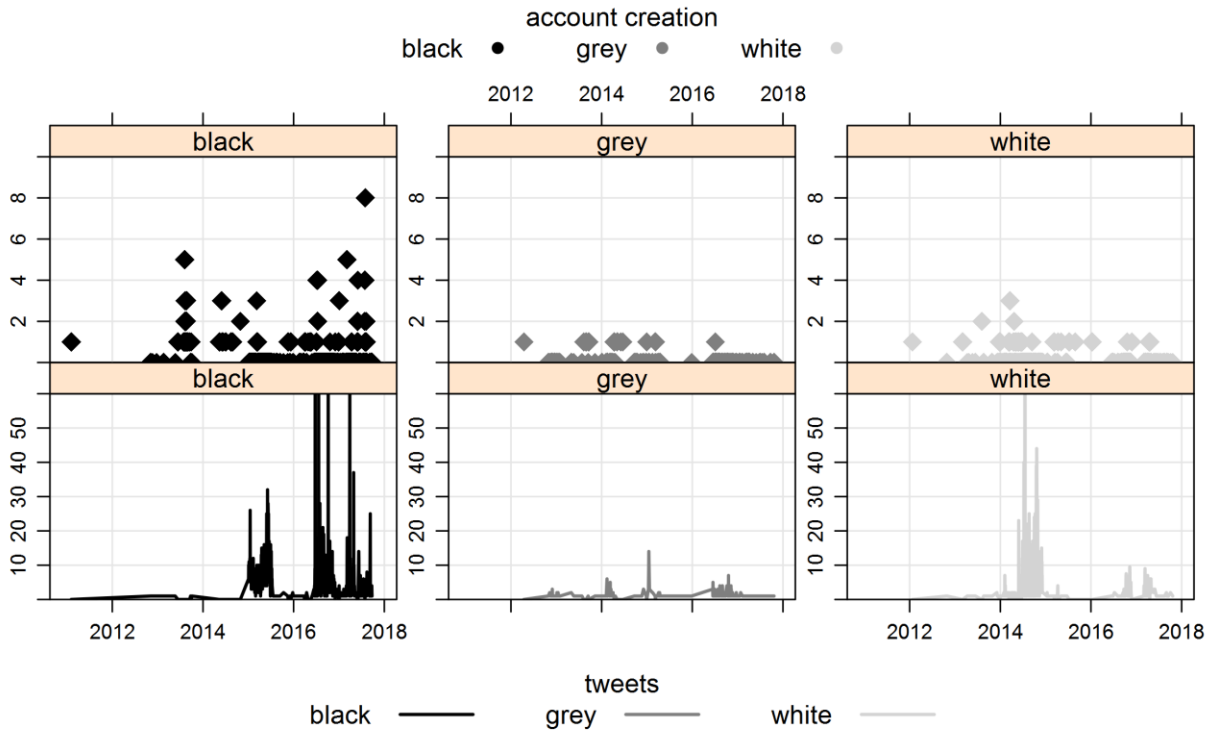

Figure 3: Account creation date and account activity for black, grey, and white propaganda accounts

Figure 3 shows that 2013 marks the inception of the black propaganda operation, with over one quarter of such accounts created in this period. These accounts however remained largely dormant until 2015 and 2016, the period when 80% of their tweets were posted. A significant uptake in the creation of black propaganda is observed in the following year (2017), but their activity decreases likely due to Twitter terminating this network of black propaganda accounts. Grey propaganda accounts, on the other hand, appear to be the most complex operation carried out by the IRA. One-third of these accounts was created in 2013 and a further 42% in 2014. While 83% of grey accounts were created before 2014, they remained largely dormant until 2016, when half of the messages tweeted by these accounts are posted. Indeed, the median activity of grey accounts falls on June 29, 2016 which is just one week after the United Kingdom EU membership referendum and right in the run-up to the 2016 U.S. elections that elected Donald Trump.

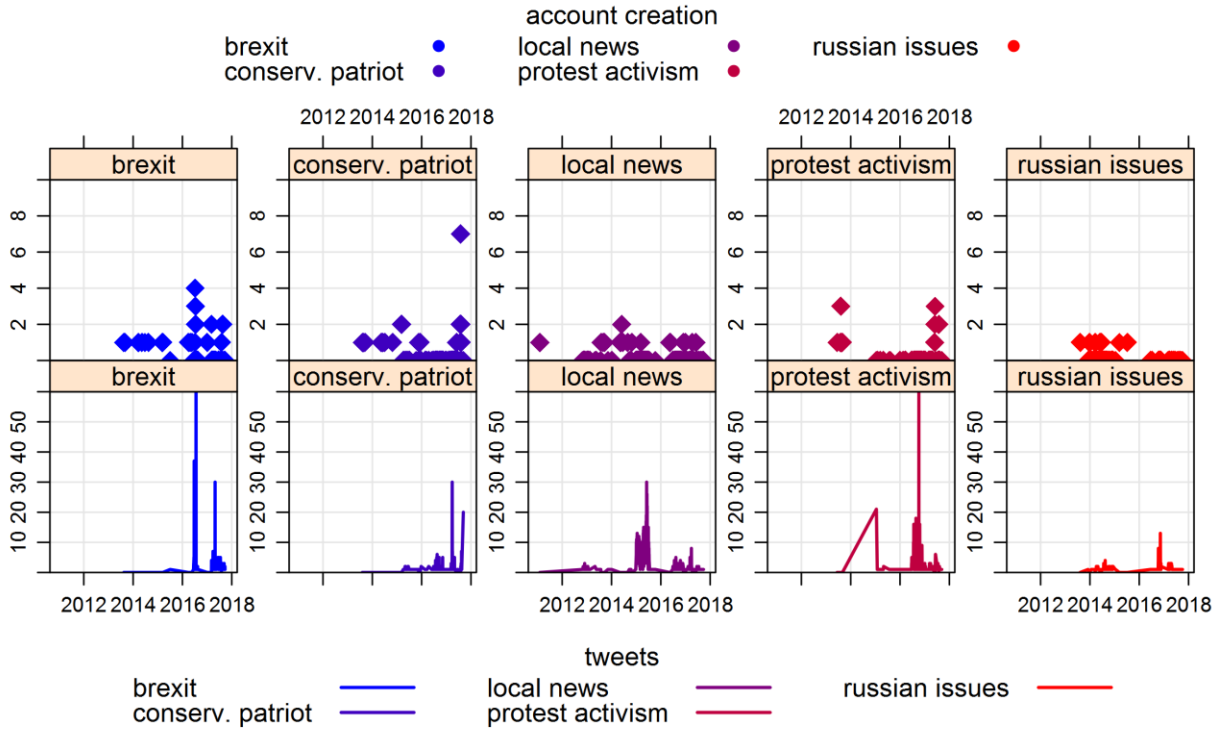

Figure 4: Account creation date and temporal patterns for IRA campaign targets

The temporal patterns identified across operations are consistent with the strategic objectives of the campaign, which can be divided in short-term, medium-term, and long-term propaganda campaigns. Short-term campaigns are often dedicated to domestic issues. Twitter accounts

covering Russian and potential Ukrainian issues, particularly news and politics, were registered between 2013 and 2014 and 85% of their activity is concentrated in 2014 and 2015. A similar pattern was observed with accounts dedicated to the Brexit campaign. While a quarter of these accounts were registered between 2013 and 2015, half of them were registered only in the run-up to the 2016 Brexit campaign. Indeed, 2016 alone accounts for 84% of the activity tweeted by these accounts. Medium-term campaigns are exemplified by the network of accounts impersonating local news outlets operated by the IRA. Sixty percent of these accounts were created between 2013 and 2014, but over 85% of their tweets appeared only in 2015.

It is however the more targeted campaigns, including conservative patriots and protest activism focusing on the Black Lives Matter movement, that display more sparse patterns of account creation followed by intense activity, likely a result of IRA securing a supply of accounts that are purposed and repurposed for targeted campaigns. Conservative patriot accounts were steadily created as far back as 2013 (21%) and 2014 (16%), but they only become active and operational in 2016, when 38% of their messages were registered, and in 2017, when 54% of this content appeared on Twitter. A similar pattern is revealed with protest activist accounts, which were largely created in 2013 when 62% of these accounts were registered, but that were only activated in 2017 when 84% of their tweets appeared. For this cohort of accounts, the lag between account creation date and activation is of nearly three years, which is a considerable departure from short-term campaigns in which accounts are created and deployed within the span of a single year. Figure 4 details the temporal differences observed across campaigns.

Discussion

The classification of IRA accounts shows that the agency deploys campaigns tailored to specific propaganda efforts, with little overlap across strategic operations. We identified nine propaganda targets with the most prominent being conservative patriots (*n*=75), black lives matter activists (*n*=50), and local news outlets (*n*=37). Common to these three propaganda targets is the use of US as self-reported location and their tweeting in English, but the hashtags used by these accounts follow a strict political agenda defined by the campaign. The other six campaigns identified in our inductive classification include Republican Community, Black Lives Matter Community, Anti-Trump Journalists, LGBT Communities; Satirical Content; and Warfare News. Figure 5 shows the three most prominent campaign targets identified in the data.

Figure 5: Account profiles of (a) Conservative Patriots; (b) black lives matter activists; and (c) local news outlets

Conservative patriot accounts claim to be US citizens and conservatives. They are self-described Christian patriots, supporters of the Republican party and of presidential candidate Donald Trump. These accounts tweeted predominantly about US politics, conservative values such as gun rights, national identity, and the military, along with a relentless agenda against abortion rights, "political correctness," the Democratic party, presidential candidate Hillary Clinton, and the mainstream media. The user shown in Figure 5a is one such example claiming to be a white male based in Texas. The profile description includes hashtags #2A (i.e., second amendment) and #tcot (i.e., top conservatives on Twitter) and amassed a total of 41,900 followers. The following tweets exemplify the topical focus of this portion of IRA accounts.

> *It's Election Day. Rip america. #HillaryForPrison2016 #TrumpForPresident*
>
> @archieolivers, 11 August 2016

> *THE SECOND AMENDMENT IS MY GUN PERMIT. ISSUE DATE: 12/15/1791 EXPIRATION DATE: NONE #VETS #NRA #CCOT #TCOT #GOP*
>
> @Pati_cooper, 18 August 2016

Black Lives Matter activists claim to be African American citizens supporting or participating in the Black Lives Matter movement. These accounts tweeted predominantly about US politics along with issues surrounding racial inequality and relied on a range of hashtags, including #BlackLivesMatter, #BLM, #WokeAF, and #BlackToLive. The account shown in Figure 5(b), with 24,200 followers, exemplifies this target of the IRA campaign. Key objectives of this effort appear to have been discouraging African Americans from voting for Hillary Clinton or discouraging voting altogether, as exemplified in the following tweets.

RT @TheFinalCall: Hillary Clinton and Donald Trump: Which one is worse: Lucifer, Satan, or The Devil?

@adrgreerr, 6 October 2016

RT @HappilyNappily: B Clinton Mobilized a army to swell jails with black bodies, Hillary led an attack on Libya, they exploited Haiti.

@claypaigeboo, 6 October 2016

The network of accounts impersonating local news outlets is the third largest propaganda effort led by the IRA. This initiative builds on the growing distrust in mainstream media and the comparatively higher trust in the local press (Newman, et al., 2016). The US branch of the campaign operated accounts that included city names and the words daily, news, post, or voice (e.g., *DailyLosAngeles, ChicagoDailyNew, DailySanFran, DailySanDiego, KansasDailyNews, and DetroitDailyNew*). This campaign also targeted German news outlets, where the IRA replicated the pattern of using city names followed by the term "Bote," meaning messenger or herald (e.g., *FrankfurtBote, HamburgBote,* and *Stuttgart_Bote*). Upon probing the data, we found they relay information sourced from established news outlets in the area they operate. The tweeting pattern comprises a single headline and does not always include a link to the original source.

When available, we resolved the shortened URLs embedded to tweets to identify the news source tweeted by disguised local news accounts. *LAOnlineDaily* tweeted exclusively Los Angeles Times content and *ChicagoDailyNew* follows a similar pattern having tweeted content from the Chicago Tribune. As such, this cohort of news repeaters seems dedicate to replicating local news content with a bias towards news items in the crime section and issues surrounding public safety, a pattern that was identified with the high scores of emotion-charge and polarization associated with the content they selected and relayed. These accounts were created between 2014 and 2017 and tweeted on average 30,380 messages per account, thus totaling over 1M for the entire cohort. They also managed to garner an average of 9753 followers per account while following only 7849, an indication that the IRA propaganda efforts might have achieved capillarity into communities of users.

Negative and contentious narratives that amplify concerns about public security, particularly crime incidents, but also fatal accidents and natural disasters, dominate the local news stories distributed by IRA posing as local news outlets. The most prolific account in our

dataset is user 2624554209 with a total of 1212 tweets. This account operated under the handle *DailyLosAngeles* in 2016, but it was also active in 2015 under the username *LAOnlineDaily* and it specialized in selecting news items from the Los Angeles Times that emphasized crime, casualties, and issues of public safety. In fact, *LAOnlineDaily* is significantly more likely to tweet headlines about fatalities compared with the rest of the IRA-linked accounts. For the 624 accounts analyzed in this study, on average only 1 in every 5 messages mention fatalities. In total, 14.9% of all tweets explicitly refer to events involving one or more deaths, while 9.3% refer to incidents with a risk of fatalities, such as violent crime, traffic accidents and natural disasters. In contrast, *LAOnlineDaily* mentions fatalities in every second tweet, with 27.5% of messages from this account explicitly referring to deaths and 23.8% referring to events with a risk of fatalities.

Conclusion

The results of this study lend support to the hypothesis that source classification remains a valid framework to understand IRA's social media operations, as account type is significantly associated with the dissemination of polarizing, populist, fear mongering, and conspiratorial content. Indeed, hypotheses 1-3 show that while grey propaganda is preferred to disseminate fearmongering stories, stoking populist sentiments, and encouraging hostile expression, black propaganda is central to efforts of sowing social discord in the target population. The testing of hypothesis 4, conversely, show that propaganda classes are significantly associated with campaign targets and lend support to the hypothesis that IRA operations are planned well in advance, with relational coordination between campaign target and propaganda class.

In summary, these results suggest fundamentally different operations tailored to achieve strategic outcomes. This is consistent with the temporal patterns identified across propaganda classes and campaign targets. White accounts were largely created and deployed as a reaction to the Euromaidan demonstrations and the civil unrest in Ukraine in late 2013. Contrary to our expectations, white propaganda accounts were frequently and overtly pro-Kremlin. These accounts were created between 2013 and 2014 and nearly 80% of their tweets appeared in 2014 in the wake of the annexation of Crimea. Black and grey operations also started in 2013, but these propaganda operations were only activated in 2015 and 2016, when most of this content appeared on Twitter. The Brexit campaign effort, however, seems to follow a short-term

organizational pattern similar to white propaganda, with a considerable portion of the accounts being registered only a few months from the referendum vote.

The campaign targets identified in this study cover a limited number of political issues and were designed to effect change on both ends of the political spectrum, simultaneously targeting the conservative base and Black Lives Matter activists. Conservative patriot accounts supported the presidential candidate Donald Trump as well as (white) national-conservative values. In contrast, Black Lives Matter accounts spoke against the oppression of minorities in the US and discouraged African Americans to vote in the 2016 elections. The IRA also run a campaign impersonating seemingly uncontroversial local news outlets, but at closer inspection these accounts curated headlines with a topical emphasis on crime, disorder, and concerns about public security. This pattern of activity is consistent with press reports (MacFarquhar, 2018) that evaluated IRA's systematic use of Twitter to sow discord in the US, to encourage white conservatives to vote for Donald Trump, and to discourage African Americans from voting for Hillary Clinton.

But we also found evidence that is at odds with what was reported in the press. Contrary to investigations reported in the media (Mak, 2018), the propaganda campaign focused on local news was not created to immediately pose as sources for Americans' hometown headlines. The accounts were created as far back as 2013 and while they have not spread misinformation, the tweeted headlines were curated to emphasize scaremongering among the population, including death tolls and crime stories, with the majority of headlines tweeted by *LAOnlineDaily* focusing on crime and violence in contrast to only 5% for the rest of the accounts. This pattern of account creation and activation shows that the IRA likely creates or purchases Twitter accounts in bulk, later repurposed to meet the needs of specific campaigns. Indeed, this pattern was observed not only in the network of accounts impersonating local news outlets, but also in the network that tweeted Brexit, Black Lives Matter, and content ideologically aligned with American conservatism.

References

Bastos, M. T., & Mercea, D. (2016). Serial Activists: Political Twitter beyond Influentials and the Twittertariat. *New Media & Society, 18*(10). doi: 10.1177/1461444815584764

Bastos, M. T., & Mercea, D. (2018a). Parametrizing Brexit: Mapping Twitter Political Space to Parliamentary Constituencies. *Information, Communication & Society, 21*(7), 921-939. doi: 10.1080/1369118X.2018.1433224

Bastos, M. T., & Mercea, D. (2018b). The public accountability of social platforms: lessons from a study on bots and trolls in the Brexit campaign. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. doi: 10.1098/rsta.2018.0003

Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review, 37*(1), 38-54. doi: 10.1177/0894439317734157

Bastos, M. T., & Zago, G. (2013). Tweeting News Articles: Readership and News Sections in Europe and the Americas. *SAGE Open, 3*(3). doi: 10.1177/2158244013502496

Becker, H. (1949). The nature and consequences of black propaganda. *American Sociological Review, 14*(2), 221-235.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*: Oxford University Press.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122-139.

Bertrand, N. (2017, Oct. 30, 2017). Twitter will tell Congress that Russia's election meddling was worse than we first thought. *Business Insider*.

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 US Presidential election online discussion. *First Monday, 21*(11).

Boler, M., & Nemorin, S. (2013). Dissent, Truthiness, and Skepticism in the Global Media Landscape: Twenty-First Century Propaganda in Times of War: The Oxford Handbook of Propaganda Studies.

Briant, E. L. (2015). Allies and audiences: evolving strategies in defense and intelligence propaganda. *The International Journal of Press/Politics, 20*(2), 145-165.

Bugorkova, O. (2015, 19 March 2015). Ukraine conflict: Inside Russia’s "Kremlin troll army.". *BBC News*.

Castells, M. (2012). *Networks of Outrage and Hope: Social Movements in the Internet Age*. Cambridge: Polity Press.

Cunningham, S. B. (2002). *The idea of propaganda: A reconstruction*. Westport, CT: Greenwood Publishing Group.

Daniels, J. (2009). Cloaked websites: propaganda, cyber-racism and epistemology in the digital era. *New Media & Society, 11*(5), 659-683.

Doherty, M. (1994). Black Propaganda by Radio: the German Concordia broadcasts to Britain 1940–1941. *Historical Journal of Film, Radio and Television, 14*(2), 167-197.

Ellul, J. (1965). *Propaganda: the formation of men's attitudes*: Knopf.

Farkas, J., & Neumayer, C. (2018). Disguised propaganda from digital to social media. In J. Hunsinger, L. Klastrup & M. M. Allen (Eds.), *Second International Handbook of Internet Research* (pp. 1-17). New York Springer.

Farkas, J., Schou, J., & Neumayer, C. (2018). Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. *New Media & Society*, *20*(5), 1461444817707759. doi: doi:10.1177/1461444817707759

Fiegerman, S., & Byers, D. (2017). Facebook, Twitter, Google defend their role in election, *CNN*. Retrieved from http://money.cnn.com/2017/10/31/media/facebook-twitter-google-congress/index.html

Gadde, V., & Roth, Y. (2018). Enabling further research of information operations on Twitter. Retrieved 17 October 2018, from Twitter, Inc. https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html

Giles, K. (2016). The Next Phase of Russian Information Warfare. In K. T. N. P. o. R. I. W. Giles (Ed.), *NATO StratCom COE* (Vol. 20): NATO Strategic Communications Centre of Excellence.

Hermans, F., Klerkx, L., & Roep, D. (2015). Structural Conditions for Collaboration and Learning in Innovation Networks: Using an Innovation System Performance Lens to Analyse Agricultural Knowledge Systems. *The Journal of Agricultural Education and Extension, 21*(1), 35-54. doi: 10.1080/1389224X.2014.991113

Hern, A. (2017). Russian troll factories: researchers damn Twitter's refusal to share data, *The Guardian*. Retrieved from https://www.theguardian.com/world/2017/nov/15/russian-troll-factories-researchers-damn-twitters-refusal-to-share-data

Hollander, G. D. (1972). *Soviet political indoctrination: Developments in mass media and propaganda since Stalin*: Praeger Publishers.

Jowett, G. S., & O'Donnell, V. (2014). *Propaganda & persuasion*: Sage.

Khamis, S., Gold, P. B., & Vaughn, K. (2013). Propaganda in Egypt and Syria's Cyberwars: Contexts, Actors, Tools, and Tactics. *The Oxford handbook of propaganda studies. New York: Oxford University Press. Google Scholar*.

King, G., Pan, J., & Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review, 111*(3), 484-501.

Linebarger, P. (1948). Psychological warfare. *Infantry Journal Press*.

MacFarquhar, N. (2018, February 16, 2018). Yevgeny Prigozhin, Russian Oligarch Indicted by U.S., Is Known as "Putin"s Cook", *The New York Times*. Retrieved from https://www.nytimes.com/2018/02/16/world/europe/prigozhin-russia-indictment-mueller.html

Mak, T. (2018, July 12, 2018). Russian Influence Campaign Sought To Exploit Americans' Trust In Local News. *NPR*.

Matsakis, L. (2017). Twitter Told Congress This Random American Is a Russian Propaganda Troll, *Motherboard*. Retrieved from https://motherboard.vice.com/en_us/article/8x5mma/twitter-told-congress-this-random-american-is-a-russian-propaganda-troll

McAndrew, F. T. (2017). *The SAGE Encyclopedia of War: Social Science Perspectives*. Thousand Oaks, CA: SAGE.

Mudde, C. (2004). The Populist Zeitgeist. *Government and Opposition, 39*(4), 541-563. doi: 10.1111/j.1477-7053.2004.00135.x

Newman, N., Richard Fletcher, Levy, D. A. L., & Nielsen, R. K. (2016). Reuters Institute Digital News Report 2016. Oxford: Reuters Institute for the Study of Journalism.

Popken, B. (2018, February 14, 2018). Twitter deleted 200,000 Russian troll tweets. Read them here, from https://www.nbcnews.com/tech/social-media/now-available-more-200-000-deleted-russian-troll-tweets-n844731

Seddon, M. (2014). Documents Show How Russia's Troll Army Hit America Retrieved 3 June 2018, from https://www.buzzfeed.com/maxseddon/documents-show-how-russias-troll-army-hit-america

Taylor, P. M. (2003). *Munitions of the Mind: A history of propaganda from the ancient world to the present era*. Manchester: Manchester University Press.

The Economist. (2018, Feb 22, 2018). Russian disinformation distorts American and European democracy. *The Economist*.

Twitter Privacy Policy (2018a).

Twitter. (2018b). Update on Twitter's Review of the 2016 U.S. Election. In Twitter Public Policy (Ed.), *Global Public Policy*.

Testimony of Sean J. Edgett, Acting General Counsel, Twitter, Inc., to the United States Senate Committee on the Judiciary, Subcommittee on Crime and Terrorism (2017).

United States of America versus Internet Research Agency LLC, Case 1:18-cr-00032-DLFFiled C.F.R. (2018).

Welch, D. (2013). *Propaganda, power and persuasion: From World War I to wikileaks*. London: I.B.Tauris.

Winseck, D. (2008). Information Operations 'Blowback': Communication, Propaganda and Surveillance in the Global War on Terrorism. *International Communication Gazette, 70*(6), 419-441. doi: 10.1177/1748048508096141

Woolley, S. C., & Howard, P. N. (2019). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford University Press.

Youmans, W. L., & York, J. C. (2012). Social media and the activist toolkit: User agreements, corporate interests, and the information infrastructure of modern social movements. *Journal of Communication, 62*(2), 315-329.