Rally Around the Tweet?

The Russian State's Use of Twitter During the Ukraine Crisis

Sean Norton

November 4, 2020

Much ink has been spilled in both the academic literature in the popular press about the Russian disinformation campaign intended to influence the 2016 US presidential election. Comparatively little attention has been paid to the Russian internally-focused information campaign, which sprung into action during the 2014 Ukraine crisis and utilized more Twitter accounts and created far more content than the subsequent US campaign (Zannettou et al., 2019). While the use of social media disinformation as a tool of foreign policy still remains under-researched and under-theorized, research on the use of disinformation as a tool of authoritarian domestic policy remains thin, despite a rapid increase in the capacity of authoritarian states to create and spread disinformation (Guess and Lyons, 2020; Bradshaw and Howard, 2018).

The existing literature on authoritarian use of social media focuses nearly exclusively on its use to censor information, a familiar topic that takes on a new twist in the internet age. As Roberts (2018) argues, overt online censorship is difficult for two reasons: the real cost and the political cost to the state. While blocking sites or deleting content online is possible, the decentralized nature of the internet and the constantly evolving methods of evading censorship turn this into a costly game of online cat and mouse. Secondly, overt censorship risks backlash against the state and even the amplification of the information that the state wished to censor. Even in China, famous for its "Great Firewall", few sites are outright blocked and little critical content is actively deleted (King, Pan and Roberts, 2013; Roberts, 2018). A more modern and effective form of censorship is to increase the cost of finding undesirable information; throttling rather than banning

websites and flooding social media with distracting news.

In contrast, I find that in the context of the 2014 Ukraine Crisis, the Russian state used Twitter not primarily to deflect or distract opposition, but rather to further push the official, state-sponsored narrative on Russian involvement in Ukraine. Rather than a unique social media strategy, this was one prong in a coordinated multimedia campaign. However, there is a key difference between the traditional media and social media campaigns: through the use of fake Twitter accounts, the Russian state effectively obscured itself as the source of the narrative, creating the impression of grassroots online support for the state's preferred narrative while also attempting to marginalize competing narratives. The effects of combining traditional media propaganda campaigns with social media disinformation are unclear, and it represents an interesting and understudied tool in the modern authoritarian propaganda toolbox.

The characteristics of the Russian domestic misinformation campaign have several interesting implications for existing literatures and raise many potential questions for further research. Firstly, the use of social media to astroturf support for a coordinated, multi-media government narrative surrounding dramatic political events provides an interesting contrast to the Chinese case, which so far has dominated the literature on authoritarian states' usage of social media disinformation. This suggests a strong need for more comparative research on authoritarian use of disinformation. Secondly, state usage of social media disinformation is not limited to authoritarian states; politicians and political parties in democracies have likewise been caught using bots and fake accounts to spread misinformation, or in some cases have directly used their own social media accounts to broadcast misinformation (Guess and Lyons, 2020). Insights from authoritarian states' weaponization of social media may be uncomfortably applicable to the study of modern democracies. In regards to the Russian case, I offer some insight as to how the Russian state's propaganda machine has adapted to the rise of new communication technologies. More generally, I raise a number of questions as to if, when, and how these campaigns are effective. While these campaigns remain small and reach only a small number of users at present, understanding when, why and on whom they are likely

to work is crucial for understanding their potential to have an impact in the future.

Social Media and Authoritarianism

Early optimism on the "democraticness" of the internet and social media has since given way to recognition that the internet, and social media in particular, represent a powerful tool for democratic and anti-democratic forces alike (Diamond, 2010; Tucker et al., 2017). Here, I outline how the authoritarian internet playbook has shifted from costly, risky, and ineffective outright censorship to softer forms of censorship, including censorship via state-funded astroturfing on social media.

The internet has proven to be a powerful tool for organizing political action, and in particular the type of broad-based contentious action that particularly frightens and endangers authoritarian regimes (González-Bailón et al., 2011; Kharroub and Bas, 2016; Bennett and Segerberg, 2012; Jost et al., 2018). It has also proved more difficult to control than traditional forms of communication. In particular, communication via social media, from Twitter to encrypted messenger apps like Telegram, is fast, can spread broadly, and is difficult or impossible to monitor. As a tool for the opposition, this allows both opposition activists and supporters to spread information on government abuses, protest, etc. quickly and at low cost on public forums, effectively circumventing government control of traditional media outlets (González-Bailón et al., 2011; Kharroub and Bas, 2016). On private social media, such as Telegram channels and WhatsApp groups, opposition activists can coordinate even when their internet or cellular connections are compromised by utilizing free end-to-end encryption. This has rationally lead to many regimes choosing a strategy of "hard" censorship, in which particularly dangerous or hard to monitor platforms are blocked, posts are removed from what social media is accessible, and authors of offending content are prosecuted (Tucker et al., 2017). In dramatic cases, authoritarians have shut down access to the internet entirely.

These hard forms of censorship are difficult to maintain in era of the modern internet. Arrests and persecution of citizens for their use of the internet can be very visible,

and while intended to create fear, carry the same potential of backlash that all forms of coercion carry. The sudden blocking of a popular site can likewise create backlash, even leading previously apolitical people to be exposed to previously invisible political information; as Hobbs and Roberts (2018) detail, the blocking of Instagram on China led to habitual Instagram users acquiring virtual private networks, allowing them to evade previously unknown or unnoticed site blocks, and exposing them to political conversation that would be censored on Chinese platforms. Additionally, as both the Instagram example and Russia's multi-year losing battle to block the encrypted messaging app Telegram illustrate, it is difficult and costly to maintain blocks, and nearly impossible to prevent blocks from being circumvented by technologically savvy users. An "internet kill switch" stands as a particularly risky strategy; completely cutting access to internet is both highly disruptive and noticeable, leading to a high potential for backlash. Additionally, complete blockades of internet access are only technologically feasible where internet infrastructure is relatively concentrated across a few links and/or a few internet service providers.

However, the decentralized nature of the internet, and social media in particular also adds a number of "soft" censorship strategies to the authoritarian playbook: rallying supporters, flooding, and friction (Tucker et al., 2017; Roberts, 2018). While pro-democracy forces can and have harnessed the "connective action" of the internet to organize against authoritarianism, authoritarians have proven effective at using social media to counter those campaigns and advance their own goals. Supporters of authoritarian rule are also prevalent on social media, and either of their own volition or by the encouragement of bot or paid poster driven campaigns, can create coordinated campaigns to attack opposition leaders and protesters or organize their own counter-protests, online and offline (Tucker et al., 2017). Likewise, the state can strategically drown out unwanted content on social media, in effect using the sheer volume of information on social media platforms to its advantage. As Roberts (2018) and King, Pan and Roberts (2013) detail, the Chinese state floods social media with irrelevant or nonsensical posts at particularly contentious moments, such as the anniversary of protest-triggering events or the early stages of potentially dangerous broad-based protest. The potentially dangerous information remains

on social media, but it becomes much more difficult for users to find the signal in the noise. Crucially, those who do not purposefully seek out the dangerous information may never be exposed to it, breaking the fragile transmission chains that underlie the explosive spread of information on social media (Barberá et al., 2015). Friction works in a similar way, but rather than making unwanted information hard to find, it makes it more difficult to access. Rather than outright banning certain content, friction involves making accessing it annoying. Throttling connections to certain sites in order to increase load times is a prominent example of this; slow page loads seem organic, and frustrate many users into giving up on accessing certain sites or content (Roberts, 2018).

These censorship strategies can be accomplished through the use of both bots and paid users. Bots are particularly useful for flooding social media quickly with specific posts, promoting hashtags, or cluttering existing hashtags (Stukal et al., 2017). However, bots are easily detectable and frequently banned by social media platforms, largely because they are incapable of more than simple interactions. Paid posters, using either their own existing accounts or running fake accounts (so-called "sockpuppets"), are more capable of evading detection and can carry out more complex tasks. Whether paid posters use their own personal accounts or create sockpuppets, the key feature is the laundering of the source, with the poster lending their credibility as a "real", unaffiliated person to the posts. This can increase the credibility or spread of pro-regime information, give the appearance of grassroots support or disapproval of a political actor or policy, or simply deflect attention away from controversial topics via flooding. In this sense, bots and paid posters both act as a form of soft censorship, boosting preferred narratives at the expense of critical narratives or simply making it harder to find critical narratives (King, Pan and Roberts, 2017).

This literature portrays social media as both an opportunity for those who oppose authoritarian rule and tool for authoritarians to blunt that opportunity and silence dissidents. While in social media's early days opposition actors were able to use it to quickly organize and broadcast decentralized protests, leading to the dramatic downfall of the Egyptian regime, authoritarian regimes are increasingly capable of sophisticated coun-

termeasures, and have begun to use social media effectively to their own political ends.

The Russian Spring and the Internet Research Agency

Modern Russia represents an interesting case for authoritarian use of social media given the reliance of Vladimir Putin on his personal approval to maintain control. Putin's most dramatic rise in popularity and highest approval ratings came with the annexation of Crimea and beginning of the Russian-supported war in the Donbass in March 2014. This was pitched to a receptive public as Russia defending fellow Russians stranded across arbitrary post-Soviet borders and simultaneously reclaiming its space as a great power on the world stage. Buoyed by state media, the annexation of Crimea became a moment of collective effervescence, where public opinion broadly and quickly unified in support of Russia's actions in Ukraine and the man seen as responsible for them, Vladimir Putin. Crimea, until that point largely forgotten to all but the most ardent nationalists, suddenly became a symbol of the ancient Russian nation and Russia's renewed greatness, and the never officially acknowledged Russian support of rebels in the Donbass became a just war (Greene and Robertson, 2018). This sudden collective outburst of nationalism, the so-called Russian Spring, played out on social media as well; nationalist groups reporting on goings-on in Ukraine and organizing support for separatist held regions sprung up on Russian-speaking social media and quickly accumulated followers and participants.

Seemingly not coincidentally, this is when the majority of Russian state-affiliated Twitter accounts first sprung into action (Figure 1), writing almost entirely in Russian (Zannettou et al., 2019). These accounts were run out of the Internet Research Agency (IRA), a pseudo-private company with ties to the Russian state and oligarch Yevgeny Prigozhin, a member of Putin's inner circle. While the Russian government has denied any ties to the Internet Research Agency, investigations led by the US Senate and federal prosecutors have established clear ties between the IRA and the Russian state (Bastos and Farkas, 2019). Leaks from inside the organization, including interviews with former employees, indicate that this was a well-organized multi-platform campaign, with targets ranging

from the comments section of online news sites to Facebook and VKontakte [1]. On Twitter, employees were expected to maintain around ten accounts which posted 50 tweets a day, with managers regularly providing employees with themes, keywords, and hashtags (Dawson and Innes, 2019; Bastos and Farkas, 2019; Seddon, 2014).

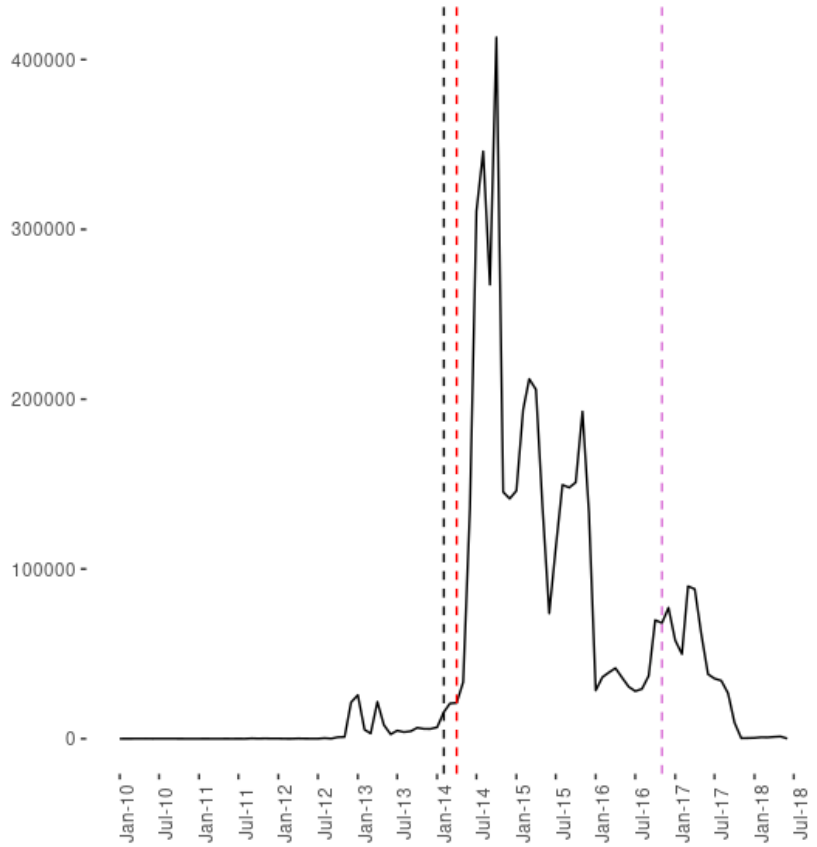


Figure 1: Russian-language IRA tweets by month. The black vertical line corresponds to the annexation of Crimea, the red line to the first sanctions imposed on Russia in response to the Ukraine crisis, and the purple line to the 2016 US presidential election.

The typical authoritarian social media playbook of censorship, flooding, friction, and harassment makes little sense in this context. Support for the Russian regime had never been higher, and even Russia's famously fragmented opposition found itself further split on the issue of Ukraine. Organizing a concerted campaign in Russian across social media for distraction or harassment alone seems destructive and wasteful when collective opinion is already rapidly turning in your favor. While authoritarians are certainly capable of erring, the context suggests that significant resources were allocated for means other than

[1]The most popular social network in Russia.

simple distraction.

In fact, available evidence suggests that the Russian intention may have been precisely the opposite. In contrast to the Chinese state, where the theories of demobilization and censorship via distraction were developed, Russia is a considerably less closed authoritarian regime. While the Chinese state largely pursues a strategy of attempting to demobilize contention and thus maintain social peace, including via censorship and flooding on social media, the Russian state often actively seeks to mobilize its supporters King, Pan and Roberts (2013); Chen (2012). As an electoral authoritarian regime, the Russian state is dependent on its ability to organize supporters both electorally and on the streets, making social media a potentially powerful, low cost tool to these ends. This is particularly true in the context of the Ukraine crisis, where the Russian state actively sought to get a receptive public involved in demonstrations, petitions, and organizing in support of Russian intervention in Ukraine (Greene and Robertson, 2019). There is also extensive evidence of a traditional media campaign to push state-friendly narratives and spin the discussion on potentially harmful topics, notably sanctions (Greene and Robertson, 2019; Szostek and Hutchings, 2015). This suggests that the IRA may have been tasked to carry out a similar campaign on social media, with the goal of deliberately building support rather than distracting or disparaging opposition actors.

Data and Methods

The Internet Research Agency's use of "sockpuppets" (or false identities) put them in direct conflict with Twitter's terms of services, and as such all content posted from Internet Research Agency accounts as well as the accounts themselves were removed from Twitter in 2017. However, in 2018 Twitter made this content publicly available as part of their continuously updated Election Integrity dataset. Twitter identified these accounts using information available only to them, and not publicly released. As such, there is some possibility of error, and in the publicly available dataset the screen names, display names, and user ID numbers for accounts with less than 5,000 followers were redacted.

Through an application process, I gained access to the unredacted dataset, containing 8,798,633 tweets from 3,479 unique accounts, of which 4,853,000 are in Russian.

To determine the topics and tactics of IRA Twitter use I adopt a two-stage approach, moving from a general description of the data to an exploration of the specific narrative surrounding sanctions. First, I use a topic model to reveal both the general topics the IRA discussed as well as their trends over time. Secondly, I train word embeddings on the corpus, allowing me to recover the context in which certain words are discussed. In addition to training word embeddings for the entire corpus, I trained word embeddings by month for the most active period of Russian language tweets: February 2014 to January 2015. This allows me to track changes in the context surrounding any given word from month to month, which I use to chart the evolving IRA strategy on sanctions.

Text was cleaned in the standard fashion; first, all hyperlinks, usernames, and hashtags were removed from the text, after which all words were stemmed. I then built a custom stop word list, using a combination of several existing Russian stop word lists as well as data-specific stop words found by examining word frequencies in the corpus. This list was used to remove these words from the corpus for the topic modeling process. Stop words were not removed when fitting word embeddings, as removing stop words destroys the context surrounding individual words.

To fit the topic models, I used a distributed online LDA algorithm available in the *gensim* (Řehůřek and Sojka, 2010). To tune k, the number of topics, I fit models ranging from 10 to 100 topics. The range of 10-20 topics performed best in terms of model coherence, using the UMass metric (Mimno et al., 2011). I then manually evaluated these 10 models, and selected the 15 topic model on the basis of clear delineation between topics and clear relation of documents within topics. While LDA is a mixed-membership model, in order to track topics over time I assigned each tweet to a topic based on the topic that is most prevalent in a given document [2].

Word embeddings were trained using *word2vec*, also in the *gensim* package, with a context window of five and using the skip-gram algorithm. I manually verified word

[2] i.e. the highest value of ϕ, the distribution of topics over a document

embeddings for both the overall model and the monthly models by testing the top 25 most similar words to several commonly used words, checking that a high proportion of these similar words were conceptually similar to the test words. I also compared the full corpus skipgram embeddings to the continuous bag of words approach, and did not receive substantially different results, suggesting relative stability in the embeddings.

Results

The topic model produced remarkably coherent topics given both the size of the corpus and the limited length of the documents (140 characters). Combined with the fact that 15 topics provide the best fit on a corpus of almost 5 million documents, this is suggestive of significant messaging coordination by the IRA. For the sake of brevity, I present the five most substantively interesting topics here, which together represent approximately one-third of all Russian-language tweets. Below, I describe these topics, present the top tokens (untranslated) for each topic, and provide a highly representative tweet (translated).

(a) **Pushing the News**: News stories, largely from Russian sources, retweeted or replied to with brief commentary

- Top tokens: жител, пострада, попа, обстрел, доллар, погибл, результат, прям, взрыв, пыта
- Example: "Thoughtfully written. The Russian people are sincerely worried about the situation in Ukraine."
- 12 % of total tweets

(b) **Putin and Patriotism**: Tweets using symbols of national pride, including discussion of upcoming 2014 Victory Day, callbacks to the Soviet victory in World War Two, and pride in the annexation of Crimea. General positive news on President Putin.

- Top tokens: крым, игр, след, запрет, эксперт, миров, сильн, кита, развит, фестивал

- Example: “Don’t get it confused: the government is the government, but the state is each of us!”
- 8 % of total tweets

(c) **Personal Opinions**: Commentary on current events posted by Russian users, and retweeted by IRA accounts. In contrast to “Pushing the New”, this does not generally include links to news articles that the retweeted user is responding to or sharing.

- Top tokens: суд, иг, сша, добр, компан, боевик, трамп, удар, бывш, луганск
- Example: “And what did you expect? Ukraine to ride off into the sunset?”
- 6 % of total tweets

(d) **America**: Discussion of US politics, US sanctions, US relations with Russia.

- Top tokens: украин, готов, ес, депутат, закон, бо, действ, госдум, предлага, похож
- Example: “Why is Russia feeding the US?”, “Trump instead of Clinton! What a week!”
- 9 % of total tweets

(e) **Russia vs. the EU**: Initially against the Ukraine association agreement with the EU, becomes heavily focused on EU sanctions

- Top tokens: украин, европ, возможн, выбор, очередн, критичн, будущ, музык, объяв, парк
- Example: “Russia hasn’t even felt a thing yet, but the EU itself can’t withstand the effect of its own sanctions!”
- 7 % of total tweets

The trend over time of these topics corresponds well to events on the ground, giving further confidence in the fit of the model. Figure 2 displays what percentage of all tweets

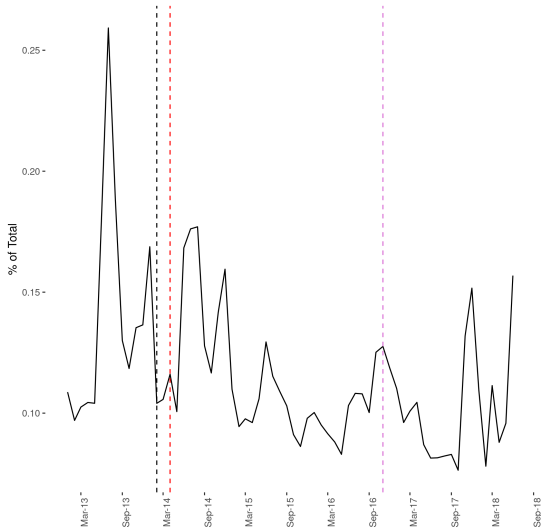
a topic represented in any given month (note that the y-axis is not on the same scale for each topic).

The changes in topic usage over time represent clear shifts in the IRA's focal points in response to events on the ground and to echo the dominant narrative on the airwaves and in the newspapers. "Pushing the News", the most prevalent of all 15 topics, experiences sustained high usage in the months following the annexation of Crimea and imposition of sanctions, before dropping off in March 2015. This fits neatly with the intensification period of the Ukrainian crisis, in which Russian state-owned/affiliated media pushed a narrative of Russians under threat from a Ukrainian government tied to far-right and neo-Nazi extremists (Szostek and Hutchings, 2015; Greene and Robertson, 2019). The IRA's intention appears to have been to amplify this campaign, with quote-tweets and comments used to give the impression of real engagement by Russian Twitter users. The "Putin and Patriotism" topic rationally experiences high usage during the period immediately after the annexation of Crimea, and also experiences its largest prevalence spike in April-May 2015, the 70th anniversary of Soviet victory in World War Two. The Russian traditional media campaign tied this anniversary to the Soviet liberation of Ukraine from Nazi rule, again emphasizing Kiev's cooperation with far-right and neo-Nazi militias on the Ukrainian front (Szostek and Hutchings, 2015; Greene and Robertson, 2019). In the America and EU topics, we see evidence of a response to sanctions, with the prevalence of each spiking not far after the initial introduction of sanctions against Russia in March 2014 and the imposition of Russian counter-sanctions on agricultural products in August 2014 [3].

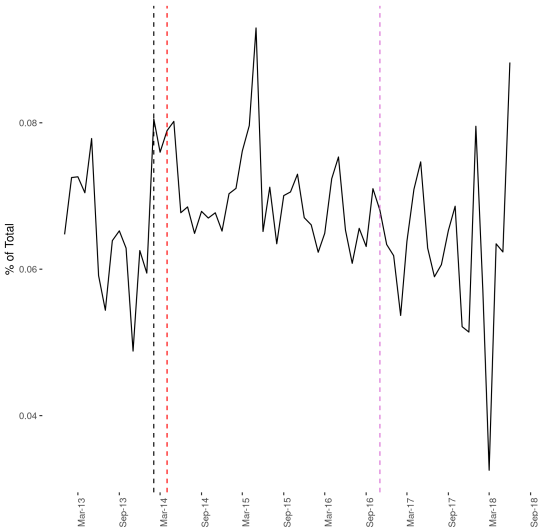
In terms of types of engagement, the dominant form of IRA engagement was low-effort retweets of already existing content, with the creators being either real users or other IRA users. However, some distinction in tactics does emerge across topics. Figure 3 shows the proportion of retweets, replies, and quote tweets across topics [4] The "Pushing the News" topic involved considerably more costly forms of engagement; while retweets still

[3]See the next section for a deep dive into the evolving content on sanctions

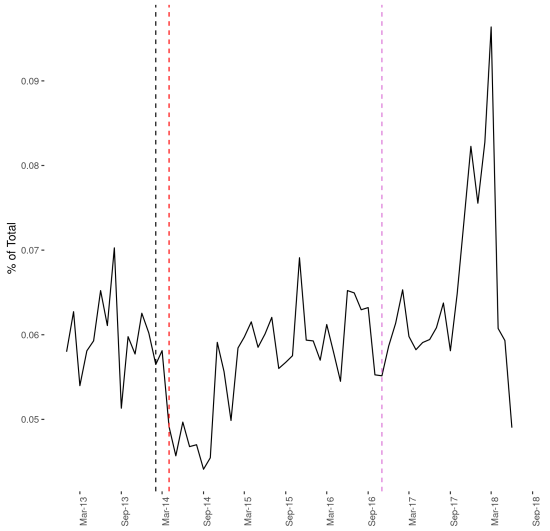
[4]This does not sum to 1, because "original" tweets are not included. Original tweets may be actually original (i.e. truly written by an IRA employee), or simply stolen from actual users or other IRA users.



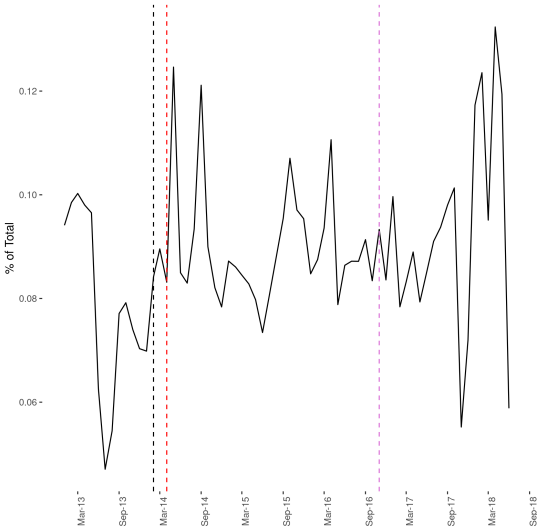
(a) News



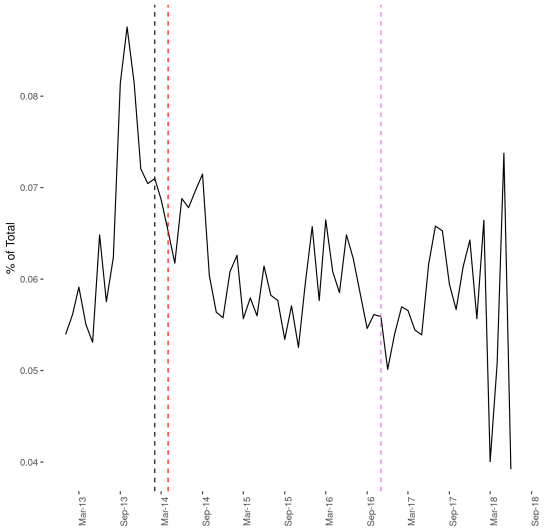
(b) Patriotism



(c) Opinions



(d) America



(e) EU

dominant, this topic features replies being used significantly more than in other topics. Interestingly, this involves more direct forms of engagement than the "Personal Opinions" topic, which appears to largely be focused on boosting the opinions of non-IRA users.

Tracking both the "Pushing the News" and "Personal Opinions" topics over time, a temporal shift in tactics also emerges. The initial post-Ukraine blitz of pushing news stories, during which this topic made up nearly every one in five tweets from April to August 2014, also represents the nadir of the IRA accounts attempting to amplify real Russians' opinions on current events. "Personal Opinions" is never a particularly prevalent topic, reaching only 7 % of monthly tweets in the period between the annexation of Crimea and the US presidential election, but it is notable that it becomes more prevalent as the attempts to directly endorse and push the Russian media narrative draw down. It is also notable that this type of platform manipulation is distinct from that of automated bots, which were extensively used over this same time period to spread news stories from Russian state-affiliated media (Stukal et al., 2017). Rather than the mindless bursts of activity designed to promote a single story/headline that are common from automated-bot accounts, serious effort appears to have been made here to present the illusion of real engagement with the story, and strategies of faking engagement shifted over time [5]. While automated bot accounts are often used in attempts to make a story artificially trend on Twitter or get picked up by many other users, the human-run IRA accounts appear to have been leveraged for a different purpose. Rather than simple amplification, the goal here appears to have been to create the perception of attention and agreement in the Russophone Twitter world. Initially, this took the form of manufacturing attention and agreement with the Russian media narrative, but later took the form of amplifying the opinions of other users that were agreeable to the Russian media narrative.

[5] Although IRA tweeters do appear to have copied each other's work frequently; many of the top tweets in this topic are either exactly the same or have only slight differences from other accounts' takes on the same story

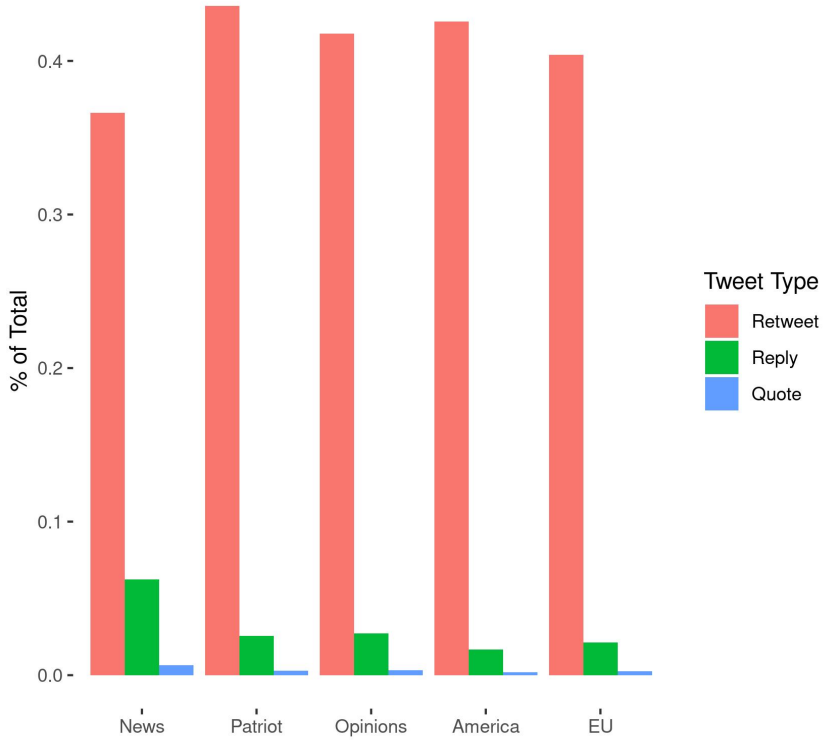


Figure 2: Tweet types by topic.

The Changing Argument on Sanctions

The IRA's attempt to create the illusion of agreement on state-friendly narratives becomes particularly clear when examining the evolving context surrounding sanctions. While the decline of oil prices in 2014 and 2015 explains much of Russia's economic woes over the period IRA Twitter accounts were most active, sanctions did have a tangible effect on the Russian economy, particularly the broad Russian counter-sanctions on food products. Sanctions also did become a salient issue for Russian citizens, with an August 2016 survey finding that 39 % of Russians were somewhat or seriously concerned with the economic fallout from sanctions (Frye, 2019).

Given the prevalent discussion of sanctions in this corpus, particularly with respects to the US and the EU, it is worth unpacking the context in which sanctions were discussed and its shifts over time. As detailed earlier, I fit word2vec models for each month from March 2014 to January 2015. For each month, I then used cosine similarity to identify the top 25 semantically-similar words to "sanctions" [6] High similarity indicates these words

[6]Since words were stemmed before being fed to the model, I calculated cosine distance to the stem

were used in a similar context with the word "sanction(s)", giving a sense of the overall narrative surrounding sanctions that is not possible with a static topic model.

The results are presented by month below. Each month includes a brief description of events surrounding the Ukraine Crisis, Russia-EU-US relations, and sanctions and the 25 most similar words to sanctions (unstemmed and translated). In some months, certain tweets were very prevalent in the discussion of sanctions, resulting in a screen name entering the most-similar words; per my data use agreement with Twitter, I do not report these [7] I also omitted the real names of people who are not prominent public figures whose tweets the IRA interacted with or named, again per the terms of my data use agreement.

- March 2014
  - *Context*: Crimea annexed, US/EU targets sanctions at specific individuals, US sanctions Bank Rossiya and forbids export of defense products. Russia sanctions specific US and EU officials in response.
  - *Similarities*: shadowing, Italian, reciprocal, introduced, perspective, against, miscalculated, criteria, isolation, anti-Russian, EU, restrictive, will introduce, tightening, extend, inflict, Europe, physical person (legal term), Russian Basketball Federation, embargo, cancel
- April 2014
  - *Context*:Situation in the Donbass rapidly escalates into armed conflict. US and EU sanction additional individuals and companies, and US restricts exports of certain dual-use goods.
  - *Similarities*: general, US, EU, IT-companies, support, closer, block, conceive, boomerang, sectoral, cheapen, introduce, skeptical, 9.02 (reference to World Bank sanction procedures), counter-sanctions, Chizhov (Russian ambassador

"санкц-", which corresponds to all declensions of the singular and plural forms of the noun "sanction".

[7] While screen names were cleaned from the corpus, I am only able to remove them when used in properly formatted replies and quotes where the screen name is preceded by an @. Due to either sloppy copy-pasting from other IRA workers or purposeful omission of the @ sign to avoid alerting the user of the discussion of their tweet, some user names remain in the cleaned data.

to EU), garden beds, EU, Russian Direct Investment Fund, agricultural products embargo, anti-Russian, avoid, Kirsan Ilyumzhinov (oligarch, head of international chess federation; ended up on sanctions lists), special measures, Yanukovych

- May 2014
  - *Context*: Donetsk and Luhansk declare independence, form unrecognized state of Novorossiya. EU targets sanctions at more individuals/companies.
  - *Similarities*: systems, presidential, introduce, conspiracy, bluff, worry, object to, anti-Russian, tightening, David Cameron, in regards to, restrictive, anti-radar, resolution, EU, diktat, embargo, symmetrical, frustrated, refer to, sanctioned (adj.), extend
- June 2014
  - *Context*: US targets sanctions at separatist commanders/political leaders.
  - *Similarities*: negotiations, dialogue, recognize, congressman, TV channel, boomerang, summit, introduce, reciprocal, Jackson-Vanik amendment, Belgrade, imposed, pressure, EU, embargo, reset, Russophobic, Albania, senseless, comical, embargo, formal, MEP (Minister of European Parliament), Iran
- July 2014
  - *Context*: Separatists accidentally shoot down MH17 over Donetsk on July 17th. Prior to MH17, the EU adds more individuals to its sanction list, and the US escalates further by sanctioning two important banks, major energy companies, and the Russian defense industry. After the shootdown of MH17, the US again escalates with broad, sectoral sanctions on the Russian energy, finance, and defense industries. The EU partially follows suit, restricting access to European capital markets and placing an embargo on arms and dual-use technology in both the defense and oil sectors.

- *Similarities*: Yanukovych, accounting, join, cross, formal, undermine, air transport, Gaddafi, sanctioned (adj.), soon, against, countermeasures, isolation, unbelievable, re-evaluation, unemployment, Kalashnikov, unpleasant, EU, restrictive, address (verb)

- August 2014
  - *Context*: US significantly restricts export of energy sector technologies. For the first time, Russia answers in a significant way: a near-complete embargo on the import of agricultural products from the US, the EU, and all other countries that had imposed sanctions against Russia.
  - *Similarities*: weapons, material, Americans, Europeans, introduced, will introduce, disastrous, pressure, counter-measures, dispute, backroom, US State Department, softened, counterproductive, against, Vygaudas Ušackas (EU ambassador to Russia), dependence, World Trade Organization, fall within, quirks, oppose,

- September 2014
  - *Context*: First Minsk Protocol ceasefire negotiated. No new sanctions.
  - *Similarities*: politics, approved, the West, organization, fellow citizens (сограждан), conspiracy, Dianne Feinstein, sanctioned (adj.), undermine, will die, energy independence, weaken, will freeze, rhetoric, embargo, US State Department, iodine deficiency, object to, push, illogical, introduce, to concern, Matteo Renzi (Italian PM)

- October 2014
  - *Context*: Widespread violations of ceasefire. No new sanctions.
  - *Similarities*: war, anonymous, politics, the West, intended, ludicrous, intrigue, veto, sadomasochism, irritated, introduce, inadvisable, contradict, against, dominate, counterproductive, dead ends, counter-sanctions, admit, anti-sanctions, fallen, hopeless, indefinitely, tolerate, parliamentarians

- November 2014
  - *Context*: 14: Minsk ceasefire falls apart. Separatists hold elections in controlled territories, in violation of Minsk Protocols. Heavy fighting resumes. No new sanctions.
  - *Similarities*: Euromaidan, pilot, actions, introduce, join, smirk, Ramzan Kadyrov, calculate, reject, the West, A Just Russia, warriors, veto, step-wise, EU, introduce, destabilization, appease, asymmetrical, unbeknownst, far-fetched, painfully, fib, agricultural products embargo
- December 2014
  - *Context*: Fighting lessens after a temporary ceasefire. No new sanctions.
  - *Similarities*: Ukrainian, partner, performing, invoke, join, effect, packet, Chizov, drifter, hostile, unlikely, valenki, extension, softened, rapprochement, undermine, symmetrical, Castro, introduce, aviation sanctions, FIBA (International Federation of Basketball), anti-Turkish, EU
- January 2015
  - *Context*: New round of Minsk talks fails to start when separatist leaders refuse to attend. US and EU ban import/export of goods/services from Crimea.
  - *Similarities*: EU, drugs (medical), declared, Rosoboronexport (state agency for export/import of defense products), introduced, packet, Ministry of Economic Development, runner, fearless, cheesemakers, Angela Merkel, softened, misunderstood, irrelevant, Luxembourg, Mohammad Zarif (Minister of Foreign Affairs - Iran), normalization, Europeans, reciprocal, diplomatic relations, introduced, Moldova, entail, absurd

Similar to the trends in topic prevalence over time, a deeper dive into the context surrounding IRA accounts' discussion of sanctions reveals clear reactions to events on the ground, as well as clear correspondence to the dominant elite and media narratives on sanctions. Initially, in March and April, the context surrounding sanctions emphasized a

narrative of miscalculation on behalf of the US and the EU, with both Russian political elites and IRA accounts prophesying that sanctions would "boomerang" back to hurt the US and the EU. Additionally, some early discussion of symmetrical Russian counter-sanctions occurred in this period, despite the official Russian position at the time being a refusal to engage in retaliatory sanctions beyond the targeting of certain officials. This narrative largely continued through April and May, but introduced more discussion of the rationale behind the sanctions and specifically the choice of targets. For example, there was considerable discussion of oligarch Kirsan Ilyumzhinov, head of the International Chess Federation, being included in the updated sanctions list. At face value, this seems absurd, though crucially the context of Ilyumzhinov's alleged financial dealings with the Assad regime were left out of or discounted in IRA tweets. In June, the discussion shifts somewhat to discontent about sanctions in the EU or states attempting to accede to the EU, notably Serbia. In July, the narrative becomes somewhat more muddled, possibly due to the reassignment of resources to the MH17 controversy; IRA tweets in English experienced a large spike in July, before disappearing until January 2015 (Zannettou et al., 2019).

In August, the much discussed potential counter-sanctions become a reality. Through September, the discussion around sanctions becomes two-pronged: one prong amplifying the justification and effectiveness of the agricultural embargo, and the other arguing that sanctions on the energy industry will hurt Europe due to its energy dependence on Russia. There is also discussion of two critics of sanctions in the EU and the US: Italian PM Matteo Renzi and US Senator Dianne Feinstein. This fits well into the overarching narrative of sanctions hurting the West as much as Russia, making them "counterproductive" and "illogical". This narrative thread becomes dominant in October, with IRA tweets decrying US and EU sanctions as "sadomasochistic" and discussing the anti-sanctions movement in the EU. From November on, the narrative becomes less clear, corresponding with both decreased discussion of the EU and US topics identified by the topic model, the start of a draw-down in Russian-language IRA Twitter activity, and several months in which no new sanctions are imposed.

As with the topic models, the close similarities between official Russian narratives and dominant Russian media narratives point to the IRA's role as a new prong in an otherwise traditional propaganda campaign. The intent, across both traditional media and the IRA campaign, was to reframe the story on the inconvenient issue of sanctions, minimizing their effect on the Russian economy while pushing negative stories about their effect on the EU's economy. The IRA attempted to amplify content in agreement with these narratives, rather than striking out to attack opponents, directly distract Russian Twitter users with irrelevant content, or engage with counterarguments to the regime-preferred narratives. The key difference, however, lies in the deliberate attempt to make this appear as real support for and interest in this narrative in the Russian Twitter-sphere.

Analysis and Further Questions

The results presented here characterize the IRA Twitter campaign as a part of a modern, multimedia propaganda campaign. Taking advantage of the relative anonymity provided by social media platforms such as Twitter, the Russian state has pursued a strategy of astro-turfing via sockpuppets, laboring to give the impression of grassroots attention and support of regime-preferred narratives. Due to the interactive nature of social media, this form of platform manipulation operates as not just an amplification of propaganda, but as a form of social censorship, potentially burying opposed narratives or enhancing the perception that these narratives are marginal.

This makes the domestic IRA campaign distinctly different from propaganda as theorized in the Cold War era, which generally divides propaganda into three categories depending on the degree of source obfuscation: white propaganda, where the source is clear, gray propaganda, where the source is not clearly identified, and black propaganda, where the source is disguised (Bastos and Farkas, 2019; Becker, 1949). While the accounts gave the impression of fake grassroots agreement with dominant Russian media narratives, seeming to fit neatly as "black propaganda", the dominance of the "Pushing the News" topic suggests that while the source of the engagement was obscured, the

source of the information was often clear. This stands in contrast to the US campaign, in which Bastos and Farkas (2019) describe a heavy focus on black propaganda, with IRA users often putting large amounts of effort into creating realistic hyper-partisan Twitter personas that spread divisive and emotionally engaging content. Fake engagement with clear sources does not fit neatly into this black/gray/white categorization. The IRA took advantage of the way social media has changed how we interact with the news by enabling instant sharing, discussion, and endorsement of stories - a level of interactivity that did not exist when the literature on traditional propaganda was defined. It simultaneously spread easily attributable narratives while obscuring the source of endorsements and opinions of those narratives, acting simultaneously as propaganda, astro-turfing, and social censorship. While the literature on propaganda remains useful for studies of online disinformation, theoretical innovation is clearly needed for the social media era (Bastos and Farkas, 2019).

However, this is not to say that the Russian strategy was effective, as Figure 3 demonstrates. Studies of the English-language campaign have found that even the millions of tweets generated by the IRA were an insignificant number of all political tweets and were seen by relatively few active Twitter users, who make up a very small slice of the total voting-age population of the United States (Guess and Lyons, 2020). Very few English-language IRA tweets received a significant amount of impressions, largely due to the fact that artificially going viral is extremely difficult; virality depends on fragile transmission chains that link highly-connected nodes in order to spread content widely, a hard phenomenon to artificially replicate (González-Bailón, 2017; Barberá et al., 2015). This is even more limiting in Russia, where Twitter is not a particularly popular social network, and where the IRA campaign's retweet counts show a similar struggle to achieve wide reach. This provides more evidence to King, Pan and Roberts (2013) and (Roberts, 2018)'s arguments that the Chinese state focuses on distraction and flooding due to their low cost; the IRA's directed and thematically-coordinated engagement appears to have generated less content to less effect (Roberts, 2018).

It is also not entirely clear why the IRA targeted so many resources at Twitter. Along

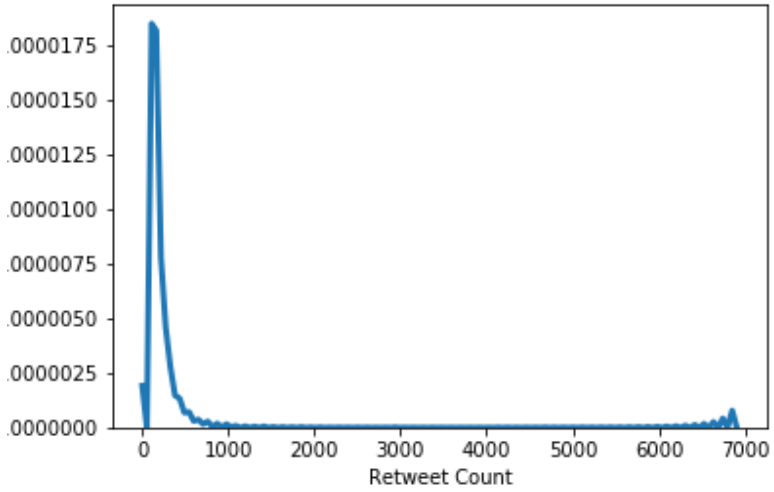


Figure 3: Density plot of retweet counts. Retweets clearly follow a power law distribution, indicating that most tweets received only one or two retweets, with a very small proportion of tweets receiving 500 or more retweets.

with Twitter not being the preferred social media platform in Russia, Russian Twitter users are more traditionally the young, educated, middle-class professionals associated with the Russian opposition. Given both the user base and the Russian state's lack of legal authority over Twitter, it has become a preferred platform for opposition activists and politicians to distribute information (Reuter and Szakonyi, 2015). Two explanations seem plausible at face value; the IRA could have been attempting to use Ukraine as a wedge issue to divide the opposition, or it may have been attempting to leverage the social censorship effect of a disinformation campaign.

Ukraine became a tricky issue for the Russian opposition, with the annexation of Crimea becoming popular even among the Kremlin-critical population, giving it great potential as a wedge issue. Existing research makes two things clear about the spread of political content, including disinformation, on social media: it is shared almost exclusively by a small set of highly partisan, high political interest users already primed to believe the content, and it is especially likely to be shared uncritically when the content inspires arousing emotions such as pride, excitement, or anger (Barberá and Rivero, 2015; Hasell and Weeks, 2016). Given the emotionally-charged politics surrounding the Russian Spring and the politically-active demographic that frequents Russian Twitter, this represented

an opportunity for the IRA to shift the salience of foreign policy to opposition-friendly voters, advantaging the Kremlin (Greene and Robertson, 2018). Unfortunately, given the retrospective nature of this study and the removal of all IRA content from Twitter, gaining leverage on this question has become impossible.

The amplification of pro-regime narratives may also have served to marginalize anti-regime narratives. Similar to the distraction described by King, Pan and Roberts (2017), fake engagement with pro-regime narratives necessarily competed with anti-regime narratives for the attention of Twitter users. Additionally, boosting the pervasiveness of pro-regime narratives gives the impression of consensus, potentially reducing the perceived veracity of opposition narratives. Of course, the key to both those effects is achieving true pervasiveness of dominant narratives, which the IRA failed to achieve.

On the other hand, this is not to say that such campaigns were not effective on those exposed or would not be effective if broad exposure were achieved. Studies of the effect of disinformation exposure on attitudes and political support are few and far between; most studies focus on either how disinformation spreads, who is vulnerable to disinformation, or the role of partisanship in the sharing and consumption of disinformation. Additionally, nearly all of these studies consider these questions only in the US context, and it is not clear that insights here generalize to other cultures or beyond the democratic context (Guess and Lyons, 2020). All existing experimental studies on the sharing and effect of disinformation take place in the context of a relatively free media environment in which facts, or at the very least alternative theories, are easier to find than in the more limited authoritarian media environment.

This leaves open a broad range of questions, which the Russian context offers some leverage on. First, comparative studies are necessary to either confirm or nuance the findings from the American case. Secondly, as an electoral autocracy, Russia presents an interesting context for the role of disinformation; the Russian state exercises considerable control over the both the traditional and online media environment, but many channels of open online expression, such as Twitter and Telegram, remain. This gives the state serious incentives to engage in disinformation campaigns in the online spaces they cannot directly

control, and to engage in different ways than more closed authoritarian regimes. Arguably, authoritarian citizens are also more likely to be experienced with disinformation than citizens of democracies, potentially making them more skeptical consumers of potential disinformation or vulnerable to different methods of disinformation. Finally, partisanship in electoral authoritarian regimes is not the partisanship of developed democracies. Who shares disinformation, why they share it, and who believes it for what reasons should be expected to differ.

References

Barberá, Pablo and Gonzalo Rivero. 2015. "Understanding the political representativeness of Twitter users." *Social Science Computer Review* 33(6):712–729.

Barberá, Pablo, Ning Wang, Richard Bonneau, John T Jost, Jonathan Nagler, Joshua Tucker and Sandra González-Bailón. 2015. "The critical periphery in the growth of social protests." *PloS one* 10(11):e0143611.

Bastos, Marco and Johan Farkas. 2019. ""Donald Trump Is My President!": The Internet Research Agency Propaganda Machine." *Social Media+ Society* 5(3):2056305119865466.

Becker, Howard. 1949. "The nature and consequences of black propaganda." *American Sociological Review* 14(2):221–235.

Bennett, W Lance and Alexandra Segerberg. 2012. "The logic of connective action: Digital media and the personalization of contentious politics." *Information, communication & society* 15(5):739–768.

Bradshaw, Samantha and Philip N Howard. 2018. "Challenging truth and trust: A global inventory of organized social media manipulation." *The Computational Propaganda Project* 1.

Chen, Xi. 2012. *Social protest and contentious authoritarianism in China*. Cambridge University Press.

Dawson, Andrew and Martin Innes. 2019. "How Russia's Internet Research Agency Built its Disinformation Campaign." *The Political Quarterly* 90(2):245–256.

Diamond, Larry. 2010. "Liberation technology." *Journal of Democracy* 21(3):69–83.

Frye, Timothy. 2019. "Economic sanctions and public opinion: Survey experiments from Russia." *Comparative Political Studies* 52(7):967–994.

González-Bailón, Sandra. 2017. *Decoding the social world: Data science and the unintended consequences of communication*. MIT Press.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero and Yamir Moreno. 2011. "The dynamics of protest recruitment through an online network." *Scientific reports* 1:197.

Greene, Samuel A and Graeme B Robertson. 2019. *Putin v. the People: The Perilous Politics of a Divided Russia*. Yale University Press.

Greene, Samuel and Graeme Robertson. 2018. "The Co-Construction of Authoritarianism: Emotional Engagement and Politics in Russia after Crimea." *Available at SSRN 3289134* .

Guess, Andrew M. and Benjamin A. Lyons. 2020. *Misinformation, Disinformation, and Online Propaganda*. SSRC Anxieties of Democracy Cambridge University Press p. 10–33.

Hasell, Ariel and Brian E Weeks. 2016. "Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media." *Human Communication Research* 42(4):641–661.

Hobbs, William R. and Margaret E. Roberts. 2018. "How Sudden Censorship Can Increase Access to Information." *American Political Science Review* 112(3):621–636.

Jost, John T, Pablo Barberá, Richard Bonneau, Melanie Langer, Megan Metzger, Jonathan Nagler, Joanna Sterling and Joshua A Tucker. 2018. "How social media facilitates political protest: Information, motivation, and social networks." *Political psychology* 39:85–118.

Kharroub, Tamara and Ozen Bas. 2016. "Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution." *New Media & Society* 18(9):1973–1992.

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* 107(2):326–343.

King, Gary, Jennifer Pan and Margaret E Roberts. 2017. "How the Chinese government fabricates social media posts for strategic distraction, not engaged argument." *American political science review* 111(3):484–501.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pp. 262–272.

Řehůřek, Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA pp. 45–50. `http://is.muni.cz/publication/884893/en`.

Reuter, Ora John and David Szakonyi. 2015. "Online social media and political awareness in authoritarian regimes." *British Journal of Political Science* 45(1):29–51.

Roberts, Margaret E. 2018. *Censored: distraction and diversion inside China's Great Firewall*. Princeton University Press.

Seddon, Max. 2014. "Documents show how Russia's troll army hit America.". `https://www.buzzfeednews.com/article/maxseddon/documents-show-how-russias-troll-army-hit-america, Accessed October 26, 2020`.

Stukal, Denis, Sergey Sanovich, Richard Bonneau and Joshua A Tucker. 2017. "Detecting bots on Russian political Twitter." *Big data* 5(4):310–324.

Szostek, Joanna and Stephen Hutchings. 2015. "Dominant narratives in Russian political and media discourse during the Ukraine crisis." *Ukraine and Russia: People, politics, propaganda and perspectives* pp. 183–196.

Tucker, Joshua A, Yannis Theocharis, Margaret E Roberts and Pablo Barberá. 2017. "From liberation to turmoil: Social media and democracy." *Journal of democracy* 28(4):46–59.

Zannettou, Savvas, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini and Jeremy Blackburn. 2019. Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science*. ACM pp. 353–362.