JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2023, 1–18
https://doi.org/10.1093/cybsec/tyad016
Research paper

Research paper

From Russia with fear: fear appeals and the patterns of cyber-enabled influence operations

Ugochukwu Etudo[1], Christopher Whyte [2,*], Victoria Yoon[1] and Niam Yaraghi[3]

[1]School of Business, Virginia Commonwealth University , 301 W Main St, Richmond, VA 23284, USA, [2]L. Douglas Wilder School of Government and Public Affairs, Virginia Commonwealth University, 923 W Franklin St, Richmond, VA 23284, USA and [3]Miami Herbert Business School, University of Miami, 5250 University Dr, Coral Gables, FL 33146, USA

*Correspondence address. L. Douglas Wilder School of Government and Public Affairs, Virginia Commonwealth University, 923 W Franklin St, Richmond, VA 23284, USA. Tel: +(571) 338-0442; E-mail: cewhyte@vcu.edu

Received 24 March 2022; revised 28 April 2023; accepted 20 June 2023

Abstract

Much research on influence operations (IO) and cyber-enabled influence operations (CEIO) rests on the assumption that state-backed digital interference attempts to generically produce sociopolitical division favorable to the perpetrator's own interests. And yet, the empirical record of malicious IO during the 2010s show that social media manipulation and messaging takes a number of forms. In this article, we survey arguments regarding the targeting tactics and techniques associated with digital age IO and suggest that existing accounts tend to ignore the strategic context of foreign interference. We propose that state-sponsored IO are not unlike conventional political messaging campaigns in that they are an evolving flow of information rooted in several key objectives and assumptions. However, the strategic position of foreign actors as an outside force constrains opportunities for effective manipulation and forces certain operational constraints that shape practice. These outside actors, generally unable to create sensation from nothing without being unveiled, rely on domestic events tied to a broad macrosocial division (e.g. an act of race violence or protest activity) to create the conditions wherein social media manipulation can be leveraged to strategic gain. Once an event occurs, belligerents tailor steps being taken to embed themselves in relevant social networks with the goal of turning that influence toward some action. We illustrate and validate this framework using the content of the Russian Federation's coordinated trolling campaign against the USA between 2015 and 2016. We deploy an empirical testing approach centered on fear appeals as a likely method for engaging foreign populations relative to some domestic triggering event and find support of our framework. Specifically, we show that while strong associations exist between Russian ad emissions on Facebook and societal unrest in the period, those relationships are not statistically causal. We find a temporal ordering of social media content that is highly suggestive of a fear appeals strategy responsive to macrosocial dividing events. Of unique interest, we also see that malware is targeted to social media populations at later stages of the fear appeal threat lifecycle, implying lessons for those specifically interested in the relationship between CEIO and disinformation tactics.

Key words: influence operations, fake news, fear appeals, BLM, machine learning, online social media, cyber operations

Introduction

Few threats to the national security and political integrity of Western democracies are as worrisome as state-sponsored efforts to poison and subvert civil society discourse through the manipulation of social media platforms and related information systems. Today, these platforms are increasingly used by nefarious actors to disrupt "information flows [via] strategic deceptions [made to] appear very credible to the people consuming them" [1].[1] The resurgence of this practice in the 21st century has been fueled by the interconnectedness borne of global Internet usage and is variously termed by pundits, scholars and others as "fake news," "democracy hacking," "election hacking," or simply influence campaigns or operations (IO) [2]. Though such interference—which sometimes happens in tandem with cyber attacks—is simply a form of political warfare given novel character by web technologies, these labels and others like them sit at the heart of both popular and scholarly commentary on the efforts of the Russian Federation, Iran, China and a growing list of other states to subvert the function of democratic civil society. This reality draws attention to such efforts, which is good. But it is also problematic because, as other researchers have noted, resulting discussion of this phenomena (1) is invariably politically charged and (2) often describes political warfare via social media as primarily consisting of the spread of strictly untrue information.

This latter suggestion—that most state-sponsored social media manipulation consists of spreading divisive untruths—is one now regularly questioned by researchers. A range of explanations for the styles of engagement seen across more than two dozen major influence campaigns against Western states since the mid-2010s have emerged, many of which offer some idea as to why untruths have received so much focus. The original argument, of course, adopted in early analyses and punditry alike, is that lies simply help inflame sociopolitical fault lines, which is the assumed goal of IO (see [3]). But this argument suffers on two fronts. First, empirical study shows that social media manipulation is clearly more than just spreading untruths (e.g. see ref. [4]). And second, assumptions about targeted division do not reflect the nuanced strategic objectives of foreign belligerents. Other explanations attempt to account for these considerations. One set of arguments, for instance, points out that the preponderance of anti-Western IO analyzed thus far are backed by the Kremlin and that interference from Russia is simply a continuation of longstanding active measures tactics [5] going back to the days of the Soviet Union. Here, the spread of lies is a more likely function of IO than other, more subtle forms of manipulation because they reflect favored standard operating procedure. Another set of arguments claims that much IO seen in the digital age thus far has been trial-and-error, with belligerents trying a wide range of tactics to see what works. In this telling, the spread of outright lies are just the most visible manifestation of multi-purpose influence efforts, the proverbial movement of the trees that reflect a hidden creature passing underneath. And yet others argue that untruthful information spread in the most prominent cases of recent IO—the Russian interference in the USA in 2016 and preceding the Brexit referendum, particularly—simply reflect case-specific knowledge (e.g. see ref. [6]). Russian efforts in 2016 may, for instance, have stemmed from generalized use of publicly available polling data or from the now-infamous "one-time trade" of campaign information by Trump associate Paul Manafort to Russian intelligence.

These explanations for the patterns scholars have now begun to observe in the prosecution of IO-related social media manipulation efforts are unsatisfying or limiting in several ways. Beyond the fact that such manipulation is obviously so often more than the clearcut spreading of lies [4, 7–12], rationalizations of IO that point to foreign state characteristics (such as institutional practice) or case details (such as the Manafort gift of information) ignore the operational realities of influence campaigns as something intended to produce strategic gain from clandestine practice. Because of this, they are difficult to generalize from and are also vulnerable to narrow counterarguments. For instance, the precedent of Moscow's old interference playbook might help explain the details of some activity in the USA in 2016, but the primary Russian institution responsible—the Main Directorate of the General Staff of the Armed Forces (or GRU)—has no clear institutional lineage to such efforts during the Cold War. And arguments about experimentation with different approaches in the 2010s have an obvious explanatory half-life, as institutions learn lessons and formalize their chosen approach at some juncture in strategic documentation, doctrine, and training. Clearly, better theorization of the common features of IO in the digital age as methods of strategic engagement is needed as both complement and contradiction to these limited or limiting arguments.

In this article, we propose that state-sponsored influence campaigns are not unlike conventional political or commercial messaging campaigns in that they are an evolving flow of information rooted in several key objectives and assumptions. However, the strategic position of foreign actors as an outside force constrains opportunities for effective manipulation and forces certain operational constraints that shape practice. In the sections below, we describe IO as emerging from a foreign state agenda that is hidden but observable. Specifically, the agenda is observable in actions that reference definable macrosocial developments in the target nation. Foreign actors, generally unable to create sensation from nothing without being unveiled, rely on domestic events tied to a broad macrosocial division (e.g. an act of race violence or protest activity) to create the conditions wherein social media manipulation can be leveraged to strategic gain. Once an event occurs, belligerents tailor steps being taken to embed themselves in relevant social networks with the goal of turning that influence toward some action. This framework constitutes a lifecycle for digital interference that is prompted by and dependent on domestic triggering events. Most importantly, this framework for understanding digital age IO is generic, allowing for explanation of IO across a diversifying spectrum of national conditions and belligerents, as well as generalization about one of the most concerning sources of insecurity for Western democracy in the 21st century.

To explore and demonstrate the validity of this framework, we perform data analyses designed to (1) surface the significance of triggering events and (2) establish the presence of a pursuant lifecycle of social media interference. On this latter point, we do so in specific reference to fear appeals [13]. This theoretical choice is intuitive, as fear appeals are arguably the most widely applied lens via which scholars in multiple disciplines have studied persuasion in the context of political campaigns (of which foreign-backed IO are one kind). Moreover, that IO involve weaponizing fear is widely acknowledged by the academic community [14–16], though a fear-centered lens has not, to our knowledge, been applied to understand the evolution of a state-sponsored information operation. Our work here thus, in addition to illustrating the significance of cycle-prompting macrosocial triggering events, also addresses a need to systematically unpack the relationship between fear appeals and IO. However, we note from the outset that our focus on fear appeals is a methodological choice. They are not critical to our theoretical contribution and other

[1] For scholarly treatments of recent information warfare efforts in strategic and operational terms, see *inter alia* Jensen 2017; Jensen, Valeriano and Maness 2019; Lukito 2019; and Bastos and Farkas 2019.

research may legitimately substitute other methods for observing IOs via indirect measurement.

To demonstrate and validate our arguments, we analyze the dynamics of the Russian State's coordinated trolling campaign against the USA beginning in 2015. We collect over 3000 individual PDF documents published by the (then) Minority in the US House Intelligence Committee—with each document corresponding to an Internet Research Agency (IRA) ad buy on Facebook—and describe the challenges in analyzing the data given its published format.[2] We note that these documents constitute the entire known population of IRA sponsored content on Facebook designed to influence the 2016 US presidential election. We then combine these data into a time series with the only publicly available collection of Black Lives Matter movement related protests, our presumed (for several reasons described below) macrosocial correlate. We design additional, exogenous covariates and fit a vector generalized linear model (VGLM) in the style of classical vector autoregressive models with exogenous variables (VARX). Using this novel dataset, we empirically test our framework and examine the strategy of IRA ad buys and the fear-based construction of IRA messaging, assessing the conventional wisdom that these buys were simplistically intended to sow discord and our alternative framework.

In simple terms, our study illustrates that coordinated digital political interference need not rely on falsehoods or so-called "fake-news" as is so often implied in punditry. In doing so, we add evidence to the work of scholars who have demonstrated that factual manipulation is often more potent than factual fabrication in attempts to set sociopolitical agendas [17]. With foreign-based influence campaigns, the ability to frame inauthentic narratives at scale toward the attainment of an unknown hidden agenda usurps organic discourse within the targeted society. While the effects of such malicious information operations are not well understood,[3] their potential to artificially promote macro-social divisions is self-evident [18]. Stewart et al. note, further, that "we have very little systematic evidence about … these accounts and how they are operated" (p. 1). This article adds such evidence by addressing the lifecycle of content publication linked to an actor agenda rather than only the content or the medium itself.

More importantly, our study validates the proposed framework of strategic IO behavior in the digital age and advances work on fear appeals by finding clear indications of threat severity, self-efficacy, and response-efficacy messaging in a major coordinated foreign state-sponsored trolling campaign. We note clear evidence that these fear appeals emissions are strongly associated with evolving ground truths; i.e. fear appeals messaging emissions are dependent on the evolution of the macrosocial divide that coordinating trolling seeks to exploit and the hidden agenda they wish to benefit. Statistically significant results tie the timing of campaign messaging to these critical triggering junctures and evidence of attempts to spread malware alongside messaging elements found later in the lifecycle strongly suggests that the sequencing that we model is purposeful on the part of foreign state belligerents.

Our work suggests a series of implications for policy and practice, not least because of the continued pressing need to better develop systems and policy regimes to combat hidden and foreign influence implemented via the manipulation of social media. The remainder of this paper is organized as follows. The "Influence operations, cyberspace, and domestic context" section provides the theoretical lens, followed by the discussion on our proposed framework and hypotheses on coordinated, non-anarchic online trolling in the "Methods: objectives, hypotheses, and a fear appeals framework for indirectly measuring IO" section. The "Data description" section presents a specific coordinated, non-anarchic activity that we chose to empirically test our framework, Russian trolling on Facebook ads, and the data collected. The "Empirical evaluation" section describes the analysis method to empirically test our framework and the test results. The "Discussion and implications: a lifecycle of IO social media engagement" section presents the discussion of results and study implications. The final section offers conclusions and future research directions.

Influence operations, cyberspace, and domestic context

The exploitation of information and information systems as a means of achieving some measure of social, economic, or political interference is far from a unique feature of the Internet age. Across human history, information warfare has been utilized to deceive opponents, to influence complex polities, and to create favorable conditions without the application of direct force. In modern history, political warfare was a significant feature of great power struggles that preceded and then defined the Second World War, as well as a core defining activity of the low-intensity contestation that constituted the Cold War. In the latter case, information became arguably the most significant weapon of both the USA and the Soviet Union short of nuclear weapons from the 1950s onwards. Over 40 years, propagandistic interference or psychological deception operations in both blocs and across the Third World played a part in almost every political transformation, economic development, and shooting war [19, 20] tied to the global contest between communism and capitalist democracy.

In the 21st century, influence operations have gained resurgent popularity as an operational mechanism for extending state power in indirect, subversive, and exploitative fashion. Unsurprisingly, scholars regularly attribute the global spread of Internet access and the evolution of web technologies as key reasons as to why IO have become so popular. Specifically, IO—for which there is an immense nomenclature of often-interchangeable terms, such as political warfare, information warfare (IW), hybrid warfare, disinformation campaigns, election hacking campaigns, and more (for a discussion of these terms, see *inter alia* [21–25])—have increasingly been deployed as one of several techno-strategic means of allowing states to degrade the power and integrity of their peer competitors without risking escalation. Cyberspace enables such operations from afar and also enables several adjuncts to traditional media manipulation tactics, such as the use of cyber attacks (i.e. cyber-enabled IO or CEIO) [26] or the facilitation of dark money activities.

Despite the growing ubiquity of influence campaigns undertaken by strategic competitors targeting one another's societal processes, IO and CEIO have yet to receive satisfying treatment by experts as a form of engagement guided and operationally shaped by clear strategic objectives and constraints [23]. As we noted in introduction, there exists a substantial body of research in communication studies, information systems, and psychology that has evaluated influence campaigns and social media manipulation, particularly taking point from major episodes like the 2016 interference in American election

[2] The data underlying this article are available at https://intelligence.house.gov/news/documentsingle.aspx?DocumentID=379. The datasets were derived from sources in the public domain.

[3] Some (e.g. Jamieson 2018) make the case, by drawing on theoretical insight but eschewing empirical analysis, that such campaigns *must* have an impact.

proceedings, British general elections both before and following the Brexit vote, and both Dutch and German federal elections between 2016 and 2017 [27, 28]. These studies are methodologically diverse and have done much to further the empirical picture of digital age IO, particularly that involving the Russian Federation. However, the question of IO targeting—meaning what drives variation in how influence campaigns are executed—remains undertheorized. Indeed, it seems fair to characterize the bulk of work on IO at this juncture, with some exceptions we will now discuss, as implicitly anchored to several core simplistic assumptions, the most obvious of which is the oft-cited but misleading idea that influence campaigns involve spreading lies in order to inspire confusion and division.

In the literature on information warfare, broadly construed, there several theories as to why digital age IO look the way they do. The variation that these explanations address is multiform, but might generally be thought of as differences in who is targeted in victim societies, how substantial is the attention paid to certain targets or narratives, and what kind of communication strategy is pursued. On this last point, while the commonplace association of the term "fake news" with IO might lead one to assume that information manipulation tends to take the form of injected falsehoods, recent analyses reveal that IO content is as varied as that coming from any political campaign [29]. Bot accounts linked to Russian campaigns, for instance, have been seen to undertake diverse messaging activities ranging from recirculating sensationalist content and implying the legitimacy of conspiracy theories to retweeting legitimate news stories and commenting on sports results [30, 31].

We group theories about the conduct of digital age IO and CEIO into four categories. The first is the most simplistic: IOs aim to create division and exacerbate fault lines of social conflict wherever they might be found. This argument is the most common one made to explain IO and is found in so many pieces of research and punditry that it might better be thought of as a foundational assumption than a well-structured theory [32–34]. The intuition behind this idea is simple: foreign aggressors generally aim to distract their competitors by inflaming domestic conflicts. This idea is, broadly, a reasonable one. But it lacks an ability to explain variation in the tactics and targets of IO and also generalizes about the goals of influence campaigns to an unreasonable degree. Some Russian-backed IO efforts have centered on efforts to discredit political process (e.g. Netherlands or Italy in 2016) while others have targeted key policies, domestic figures or even foreign policy issues (e.g. the UK in 2015–16 or Czech Republic in 2017–18) [27, 35]. The sociopolitical division assumption broadly fits the facts, but cannot explain variation in tactics or technique.

Two additional sets of explanation for digital IO emphasize opposite sides of the target/victim coin. On the one hand, analysts have pointed to Russian influence campaigns as an extension of Soviet-era employment of active measures for purposes of political interference [5, 36, 37]. Here, as noted in the sections above, it is the context of the attacker that explains variation in approach. Thus, one could extend the argument to other belligerent states. But, again, while there is likely some truth to the fact that institutional and national history drive contemporary approach, there is enough evidence of the varying and changing nature of IO in the 2010s and 2020s thus far to illustrate that Russian tactics and techniques are evolving with the technologies and socio-technological context of the 21st century, even as the strategic preference for IO remains static. On the other hand, some experts have argued that formats of IO engagement in certain cases is down to the facts of a specific case [6]. There is circumstantial evidence that in 2015–17, for instance, Russian operations in the USA and the UK benefited from information handed over by domestic surrogates. Again, there is likely some truth to these assertions. However, such arguments offer a limited ability to generalize about IO. Not all foreign belligerents can access credible domestic sources of targeting information. Not all IO campaigns favor one domestic element but rather many or none, seeking more generally instead to suppress voter turnout or to generate issue support. And even those belligerents that do use proprietary information from surrogates do not do so in all cases, as has been true for Russia across the nearly two dozen IO launched against the West by Moscow since 2014.

Finally, many have expressed the idea that much IO seen to date launched by countries including Russia, China, Venezuela, Iran, Cuba, and Saudi Arabia have contained a great amount of trial and error [38]. This argument has an obvious flaw if one were to attempt to apply it to explain variation in patterns of IO engagement, namely that experimentation will invariably lead to predictable practices for an actor after some *n* amount of time. And yet, the substance of this explanation for IO is hard to overlook. On the one hand, early efforts by Russia to manipulate social media for strategic gain in the mid-2010s did—as noted by numerous writers (e.g. [39])—appear to try different methods of engagement before settling on preferred techniques. On the other hand, more importantly, the idea of relying on conditions in the target nation to set the tempo and style of engagement makes substantial sense. After all, civil societies differ in the base conditions that might prove exploitable for foreign belligerents. Moreover, forcing a new divide or injecting sensationalist narratives without context might draw attention to outside interference and risk hardening the target nation against such influence going forward.

We also note that many studies exist that focus entirely on the IRA without theorizing more broadly on IO. Specifically, in the years following the House Select Committee on Intelligence's release of IRA ad-buy data, and coinciding with the concerted efforts of social platform owners (e.g. Twitter) to make available data related to coordinated trolling campaigns, numerous studies have emerged offering up varying analyses of that data, attempting to account for the actions of the IRA. This literature does the critical work of providing thorough characterizations of the content of IRA coordinated trolling activity. Such studies have shown that IRA coordinated trolling employed rhetorical techniques thought to elicit "anger and fear" [40]. These authors find evidence that inflammatory language was widely used in IRA advertising content on Facebook. Others have looked at how different IRA ads on Facebook were consumed by the target audience, finding that right-leaning content was most voraciously consumed [41]. The University of Oxford has published one of the most exhaustive characterizations IRA sponsored social media content [2]. However, the messaging content of the IRA social media campaigns are not within the scope of their analysis, which is focused on the structured data points provided by the social media platforms and the US Congress. Others have focused squarely on the engagement [42] of the audience with the IRA sponsored content. These authors also employ a topic model (as well as other content analytics such as sentiment modeling) to shed light on the correlation between content type and user engagement. Interestingly, we derive an almost identical topic model in the following sections of this study. What is common across every study we examined is their focus on cross-sectional analysis. No study that we are aware of have studied time-varying behavior of IRA content and, critically, we are unaware of any novel frameworks for coordinated trolling that have been validated using IRA content.

In each of the theories described above, there are nuggets of robust logic. However, each foregoing argument is limited or limiting in its articulation. Instead, we suggest here that influence operations are not dissimilar to conventional political or commercial messaging campaigns in that they are an evolving flow of information rooted in several key objectives and assumptions. That said, there is a unique feature of state-sponsored IO that sets them apart, which is captured by some of the above arguments, namely that belligerents are generally unable to generate influence from nothing due to the risks of being unmasked. This outcome is undesirable due to the costs of being exposed, including reduced receptivity to outside influence in the target state and the possibility of conventional escalation. For those attempting to explain patterns of IO, this means that the agenda of belligerents is always hidden, albeit observable in the facts of engagement spread over different levels of social media manipulation and associated activities. The outsider-looking-in dynamic also means that belligerents necessarily rely on domestic events to provide the context and initial interest in current events from which influence can be built, expanded, and leveraged to create real effects (such as voter suppression or activation). This simple framework, we argue, complements the foregoing commonplace theories about IO and presents a unifying logic of when we might expect to see distinct forms and flavors of influence campaign tactics.

Methods: objectives, hypotheses, and a fear appeals framework for indirectly measuring IO

As our focus is to unravel the construction and application of coordinated influence campaigns in relation to a specific set of triggering events, we must first develop an empirical basis for establishing when evolution of IO-linked content does occur. We do so via reference to fear appeals.

Fear appeals: structure and purpose

Our unifying framework for the narrow arguments that dominate much thinking about IO assumes that outsider belligerents *must* leverage domestic developments to gain initial traction and influence in a targeted state, after which—indeed, *only* after which—the more conventional assumption that influence campaigns build influence toward some actionable objective plays out. Therefore, evidence in support of this framework would first include changes in messaging behavior on social media platforms by a foreign interferer that is sensitive to macrosocial dividing events. Then, we would expect to observe a pursuant lifecycle of translating influence to action to reflect the hidden agenda of the foreign interloper. Our research questions are, thus:

- Does the emission of a specific IO appeal in a coordinated IO campaign react to evolving ground truths (i.e. a macrosocial dividing event)?
- Second, does a broader, suspected, hidden agenda appear to affect the emission of messaging in a coordinated IO campaign?

Significantly, the content of the foreign state's agenda is irrelevant to our testing. While such a hidden agenda is theoretically revealable via analysis of the visible manifestations of IO on social media and in other settings, our aim here is simply to demonstrate that efforts to advance some agenda is linked to domestic triggering events.

For our empirical testing strategy, we turn to fear appeals as one such theoretical framework within which to nest understanding of IO messaging. Fear appeals are persuasive messages designed to alter a target individual's (or group of individuals') behavior by arousing their fear of danger, harm, or even discomfort [43] and are quite arguably the most widely referenced lens via which experts across multiple disciplines have studied political persuasion. In no small part, this is because the idea behind the fear appeals concept is simple. A scared individual changes their conduct to address the source of their fear. And indeed, there are pre-existing links to IO research that further validate this choice as a vehicle for our empirical testing. Specifically, work on directed trolling and bot warfare, particularly Jamieson's [44] discussion of Russian interference in 2016, has suggested that an understanding the effects thereof might best emerge from the examination of coordinated trolling content emissions as fear appeals.

Naturally, fear appeals are more complicated in their application than simply the conveyance of a threat. Threat determination depends, to some degree, on the perceptions of the targeted individual (or group) over and above the objective characteristics of the threat. A distinction between the objective and subjective characteristics of threat messaging is important to understanding the impact of fear appeals because it fundamentally affects the likelihood a targeted individual will take steps to mitigate their anxiety. The other variable element that affects the arousal of fear is the potential efficacy of responses available to a targeted individual. Decades of study have shown that the manifestation of fear absent some consideration of situational reaction is unusual in humans, not least because of our in-built biological priming toward fight or flight [45]. Upon being presented with a threat, individuals must ascertain the probability that said threat can be mitigated alongside an assessment of one's own capability for taking the needed action. Assessment of possible responses, as with the threat itself, is subject to both external and internal cues. Limited self-confidence, for instance, might lead to a skewed assessment of one's ability to succeed in combating the source of one's anxiety, thus leading to otherwise unexpected inaction.

Fear appeals as a targeting strategy

A significant clarification made by scholars about the relationship between fear arousal and changes in behavior has to do with the manner in which threats to the individual prompt mitigation strategies of variable form, depending on the nature of the fear being experienced. Janis [46] famously argues that the difference between some fear arousal and too much fear arousal is the difference between the adoption of adaptive and maladaptive strategies. With the former, individuals' reactions take aim at the actual threat. With the latter, reactions emphasize the alleviation of fear itself.

Today, this important distinction in target response to such triggering messaging is best articulated in Witte's extended parallel process model [13]. Witte, building from the work of Leventhal [47], and Rogers [48, 49] on the susceptibility of audiences to fear inputs, argued that individuals arrive at one of three reactive conclusions—do nothing, fear control (maladaptive strategies) or danger control (adaptive strategies)—dependent on (1) how impactful a threat appears and (2) how effective it appears a response might be. As we describe above, the significant elements of these calculations are the interaction of personal and situational facts. Assessments of impact emerge not only from an objective notion of threat potential but also from a perception of personal risk. Response effectiveness is a question not only of rational cost–benefit analysis of steps that might be taken to mitigate risks, but also of capacity-oriented self-assessment.

Fear appeals are a convenient and intuitive methodological focus for our attempt to observe IO campaign behavior via indirect study of IO-linked content. After all, they clearly work as a targeting strategy for conventional political campaigns. Long-standing efforts to assess voter defection in swing districts have demonstrated the ability of messages that are carefully crafted to convey a threat to successfully change the composition of audience preferences, at least at the margins [43]. A seminal 1940 study of Lake Erie area voters by Columbia University [50, 51] also showed that ∼8% of

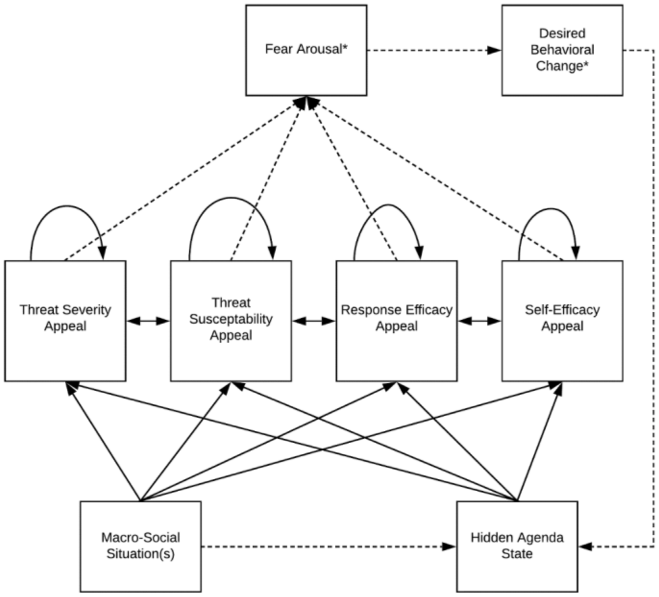


Figure 1: A fear appeals framework of coordinated trolling. *unobservable (dotted line indicates unobservable processes).

cross-party defections during that year's presidential election could be explained by exposure to agenda-setting media coverage. Significantly, that study showed that mass media content appeals often simply reinforce the inclinations of voters. But it also demonstrated that tailored messaging could prompt individuals to move away from their declared party's position by appealing to core social or political beliefs. In those cases, reinforcement occurred but underlying preferences overrode inclination toward a party.

At the level of an agenda-setting campaign, fear appeals must be targeted along two lines. First, the content must be substantively relevant to the issues and context of relevant target individuals. Second, fear appeals must simultaneously strive to articulate heightened threat severity and high levels of efficacy. Such targeting is not simple, because any attempt to persuade even a population of reasonably uniform predispositions at scale must strike a balance between customization and consistency in messaging when trying to induce fear arousal. Fortuitously, however, this provides yet additional reason as to why fear appeals present as a valid vehicle for our empirical testing. Simply put, fear appeals are a likely choice of approach of foreign belligerents because they benefit from the same conditions that outside IO must reference—an organic macrosocial development that can be leveraged to build and use influence. In the next sections, we look for and find such targeting.

Hypotheses and empirical design

With fear appeals, our expectation of threat projection pivots significantly on the notion of *distance* between the apparent threat and the targeted subjects (a factor built into the differentiation between issue-specific and context-specific messaging within the fear appeals literature). While it is logical that we might expect IO campaigns to convey concerning themes and scenarios, such content should merely constitute the jump-off point for attempts to translate influence into activity. We anticipate seeing incidences of fear appeal messaging (i.e. messaging that specifically moves beyond simple threat presentation to engage the audience with efficacy appeals or "capture" efforts) following such threat representation as representative of active IO management. Specifically, we should see evidence of general efficacy messaging proceeding such content, followed closely or contemporaneously by self-efficacy communications. However, we expect to see a clearer relationship between the execution of structured fear appeal messaging and incidence of a specific triggering event. Then, in instances where this sequence of messaging plays out, we anticipate efforts to "capture" the audience and promote the cause.

In Fig. 1, we present our fear appeals framework of coordinated online trolling. We identify the four fear appeals messaging types: threat severity, threat susceptibility, self-efficacy, and response efficacy [13] and visualize the assumed relationships between the various appeals and the environment outside of the control of the coordinated trolling campaigner.

The appeals

Threat severity and threat susceptibility messaging are important to our framework as they tie in external stimuli related to the focal macro-social situations. Threat severity appeals communicate the magnitude of the problem, and susceptibility appeals communicate the audience's vulnerability to the threat [48] posed by the macro-social situation. Efficacy messaging follows from the establishment of the threat. A response efficacy appeal both proposes some response to the threat and communicates that the response will be effective. Self-efficacy appeals communicate the ease with which a response can be carried out [47, 48].

The external environment

Our framework assumes that the macro-social environment in which the coordinated trolling operation is executed is *outside of the control of the operators* but is assumed to affect their agenda. The hidden agenda depicted in Fig. 1 is never truly known to observers of the trolling operation, but can be reliably approximated. The operators, in our model, are assumed to attempt the elicitation of some behavioral or psychological change in a target audience by attempting to arouse fear in that audience. Neither fear arousal nor behavioral change are directly observable. An appeal to fear is just that—an appeal, and "[t]rue threats do not always inspire fear and people sometimes experience fear in the absence of true threats" [52].

A coherent fear appeals strategy must be modulated and our model assumes that this modulation is a response to some proxy for the hidden agenda. A unique contribution of our approach is that it is stateful and dynamic. We understand the hidden agenda to be a state of affairs with respect to some proxy at a given time. While those state transitions are outside of the control of the coordinated trolling operators, the state of hidden agenda, we hypothesize, influences the operators' messaging.

Given our argument, the evolution of a triggering macrosocial situation should either accommodate threat and efficacy appeals or discourage their use. As we noted previously, an effective framing of world events, must be empirically credible or experientially commensurable [53]. Because threat and efficacy appeals are both environmental and messaging cues [13], their *emission* is a function of the evolving empirical reality *as well as past messaging*. To effectively arouse fear, threat and efficacy appeals need to draw upon real world events and build upon previous messaging. The hidden agenda state is, on its face, separate from the macro-social situation that coordinated trolling operations seek to manipulate. However, changes in this state over time will similarly afford opportunities for effective fear appeal emissions or preclude such emissions. We also note that while the trolls may obscure their hidden agenda, it *must* be empirically related to the focal macro-social issue. This twin dialectic of resistance and affordance [54] jointly imposed by the empirical ground truths *and* the emerging fear appeals strategy, is both observable and closely monitored by coordinated troll operators. Together and over time, these fear appeals are theorized to result in fear arousal in at least a subset of the target audience which may lead to behavioral changes. A potential causal linkage between the efficacy with which coordinated trolling output arouses fear and behavioral change in the target audience is currently not directly observable. Coordinated trolling operators, however, may view changes in the hidden agenda state as evidence of behavioral and attitudinal change in the target audience. Irrespective of any assumptions on the part of troll operators regarding the causality of message emissions with respect to this change, it remains plausible that troll operators modulate content over time subject its state [42, 55].

Efficacy appeals cannot exist independently of threat appeals. A fear appeals messaging strategy, if present in coordinated online trolling campaigns, should reflect this. In the extended parallel process model (EPPM) of fear appeals [13] the relationship between threat appeals and efficacy appeals in fear appeals messaging is emphasized. The first proposition under EPPM stipulates that high perceived threat leads to message acceptance when perceived efficacy is also high. It is important to note that levels of threat perception and efficacy perception are modulated in the message content constituent of fear appeals messaging strategy. The EPPM goes further, noting in another of its propositions that increases in threat perceptions when efficacy perceptions are low leads to overwhelming fear and threat avoidance [13]. For instance, it would be poor strategy to ask an audience to stop smoking without first framing smoking as being bad for their health. To affect behavioral change in a target audience, a coordinated online trolling campaign will need to first establish some threat, *then* motivate action by introducing efficacy appeals:

H_{1a}: Threat fear appeals will be followed by response efficacy appeals.
H_{1b}: Threat fear appeals will be followed by self-efficacy appeals.

It has been argued [56] that there is a strong relationship between what we term here as "macro-social division" (seen in Fig. 1 as "Macro-Social Situation") and fear. By macro-social division, we mean a macro-social shock that generates macro-social stress. This shock may be in the form of publicized events that heighten existing racial (e.g. intense media coverage of Black Lives Matter protests and demonstrations) and economic divisions (e.g. intense media coverage of the Occupy Wall Street Movement) in a society. By macro-social stress, we mean large scale anxiety within a society at a given moment in time driven by some fear inducing or threatening ground truth, or, "the evolving macro-social division" [57, 58]. Macro-social tensions are closely related to fear. Arousing fear in large swathes of individuals across a society fosters the development of divisive, nationalistic, and authoritarian views:

> "Fear is then a recurring social reaction when it comes to conditions of great change and uncertainty. It is possible to think that the generalization of competitive conditions encouraged by globalization and the collapse of traditional regulations produce many social situations likely to cause fearful reactions. Such reactions run parallel to the rejection of Others and their difference, since that fear of the Other is merely the expression of an individual's own fear."

We also discussed earlier how effective fear appeals are purposeful framings that require empirical commensurability to be effective. Bundling together the fear arousing nature of macro-social division, and the role of an evolving empirical reality in creating affordances for certain messaging types, we propose that the emission of message content within the framework is, at any point in time, is a function of that empirical reality.

First, we argue that the intensification of the focal macro-social division affords efficacy messaging:

H_{2a}: Response efficacy and self-efficacy fear appeals will be emitted to coincide with real-world events that intensify the focal macro-social issues either contemporaneously or shortly thereafter.

Similarly, the intensification of the focal macro-social issue affords threat appeals messaging:

H_{2b}: Threat fear appeals will be emitted to coincide with real-world events that intensify the focal macro-social issues either contemporaneously or shortly thereafter.

In addition, we address the evolving state of the hidden agenda that drives the coordinated online trolling effort. We reiterate that the hidden agenda is, on its face, unrelated to the focal macro-social issue at the heart of the fear appeals strategy. However, the state of this agenda at any given time will lead to the emission of fear appeals messaging. When the state is unfavorable (i.e. a signal to operators that the strategy is not working), we expect that both threat and efficacy messaging will be ramped up:

H_3: Unfavorable states of the hidden agenda will lead to increased threat fear appeals emissions.
H_4*:* Unfavorable states of the hidden agenda will lead to increased efficacy fear appeals emissions.

Data description

The current study benefits from actions taken by Democrats of the House Intelligence Committee in the USA who released the full text of—as well as metadata associated with–3393 promoted Facebook posts (advertisements) linked to the IRA in semi-structured PDF documents.[4] Both the authenticity and attribution of the content were thoroughly vetted by investigators, as well as Facebook. However,

[4] https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.

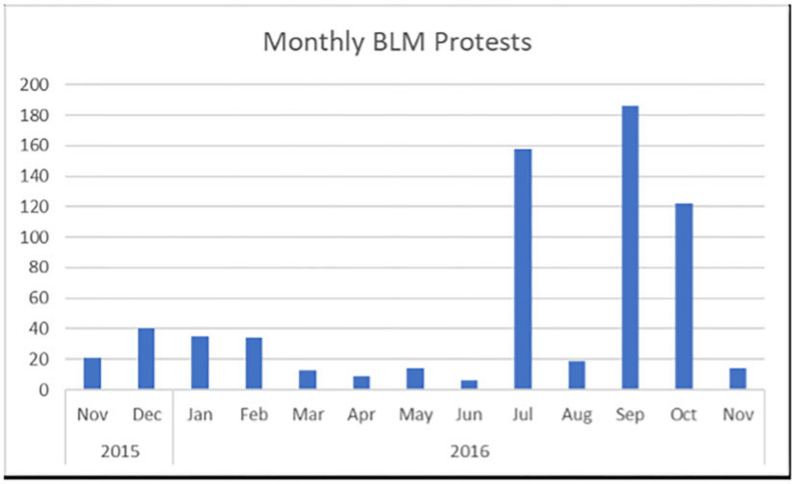


Figure 2: Monthly count of BLM protests.

there are accessibility challenges with the data release. House staff appear to have scanned printed copies into PDF format; the data are embedded with the documents as images. This poses a challenge for analysis. To resolve this challenge, we automated a pipeline—sing a Java API for Tesseract OCR (Optical Character Recognition) in concert with Oracle's PDFBox API—that traverses each PDF document and applies OCR to the primary image embedded within the first page of the document.

The documents follow a standard format where the first page contains fields regarding the content timing and other metadata germane to the advertisement. The second page is the actual rendition of the advertisement as it would have been shown to users. We do not make use of the second page. Once an image is successfully extracted from a PDF, the Tesseract OCR API is used to extract legible text. The pipeline ends with the application of heuristics to extract data from relevant fields for structured analysis. To inspect the accuracy of the OCR process, we randomly sampled 30 ads from the collection. The OCR classifier did not significantly misclassify characters in any of the sampled advertisements.

We also collected a dataset of 1921 Black Lives Matter related protests and demonstrations from August 2014 through to May 2018 as a likely source of macrosocial division observed by the IRA. This choice is also intuitive given that, as other IO researchers have noted, the Russian campaign in 2015–16 in the USA placed an unusual degree of focus on African–Americans [59]. Our BLM dataset was crowd sourced from http://elephrame.com, a site dedicated to tracking these demonstrations. We automated the collection of data from this repository and verified our collection using media links to the demonstrations provided on elephrame.com. In Fig. 2, we chart the count of BLM protests by month. The second half of 2016 saw a massive uptick in the count of protests especially in the periods prior to the November 2016 presidential election. We will explicitly account for election proximity later.

In Fig. 3, we chart the count of IRA advertisements by quarter. Initially, when comparing the chart below with the chart above, there is an unmistakable alignment between the protests and IRA ad buys on Facebook; specifically, there appears to be a lagged relationship. Below, the IRA ramped up its ad buying activity in the second half of 2016 in general, and in the fourth quarter in particular. Whereas BLM protests ramped up in the third quarter of 2016, the IRA appears to have acted in the following quarter, sustaining that action into 2017. Clearly, there is a need to systematically confirm this relationship.

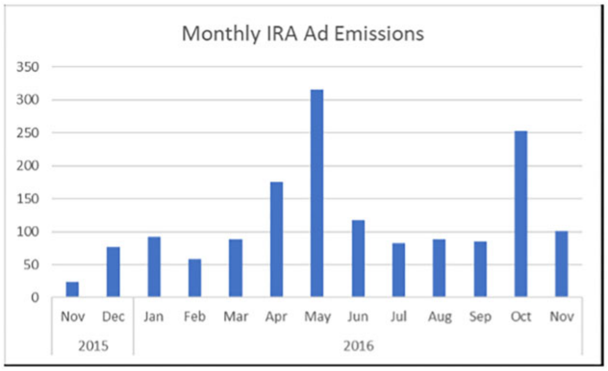


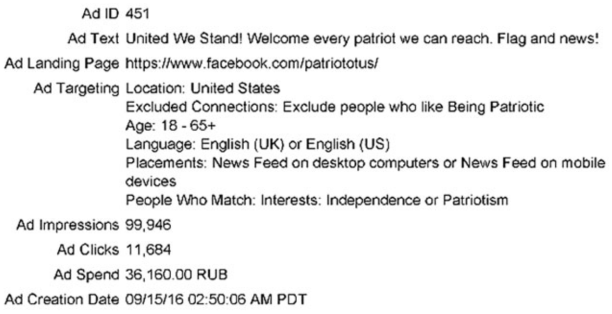
Figure 3: Monthly count of IRA ad emissions.

Each event recorded on http://elephrame.com is supported by a reputable news source and identified by a URL. As before, we randomly selected 30 protests from the dataset and found that all events in the sample were verifiable and correct. As the site did not provide a mechanism to download the collection, we automated the traversal of the site's pages to collect the comprehensive list. Once the comprehensive list was established, we merged both datasets into an SQL database for analysis.

Further, we obtain data on police shootings, another focal macrosocial issue, from the Washington Post's database on fatal police encounters.[5] This database allows us to filter fatal encounters such that we only consider fatal police shootings of African–Americans. We use the Google Trends API to obtain the daily weighted search popularity of Black Lives Matter over the relevant study period. We use FiveThirtyEight's polling data to obtain the Clinton–Trump polling spread, which we take to be the hidden influence agenda. FiveThirtyEight[6] provides a downloadable collection of reputable polls covering the relevant study period. Indeed, it is this polling data that bounds the study period. The 2016 presidential election polling data are only reliably available from 17 November 2015 until 8 November 2016—the day of the presidential election. In November and December of 2015, polling data are sparse. As such,

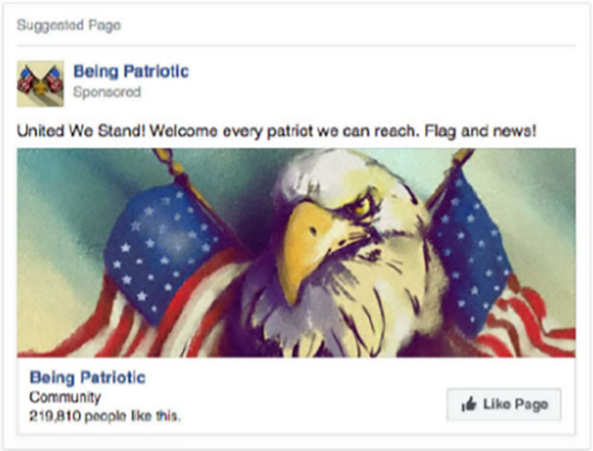
[5] https://www.washingtonpost.com/graphics/2019/national/police-shootings-2019/?noredirect=on.

[6] https://projects.fivethirtyeight.com/2016-election-forecast/.



Ad ID	451
Ad Text	United We Stand! Welcome every patriot we can reach. Flag and news!
Ad Landing Page	https://www.facebook.com/patriototus/
Ad Targeting	Location: United States Excluded Connections: Exclude people who like Being Patriotic Age: 18 - 65+ Language: English (UK) or English (US) Placements: News Feed on desktop computers or News Feed on mobile devices People Who Match: Interests: Independence or Patriotism
Ad Impressions	99,946
Ad Clicks	11,684
Ad Spend	36,160.00 RUB
Ad Creation Date	09/15/16 02:50:06 AM PDT

Figure 4: Sample page 1.



Suggested Page

Being Patriotic
Sponsored

United We Stand! Welcome every patriot we can reach. Flag and news!

Being Patriotic
Community
219,810 people like this.

Figure 5: Sample page 2.

while we require a daily time series, there are several consecutive days for which polling data are unavailable. We assume that the IRA, if it is using polling data to determine its advertising emissions, would do so based on the last available polls or on the proprietary campaign information provided by Manafort. We therefore construct a series for the Clinton–Trump polling spread that relies on the last available polls for each daily period. Finally, in Figs 4 and 5, we show two typical examples of the raw data from which we parsed out the details of IRA sponsored content on Facebook.

Empirical evaluation

We first identify the fear appeals messaging types as the necessary first step to showing a meaningful evolution of IO activity based on domestic triggering events. Then, we develop a set of empirical models to test our hypotheses.

Identifying fear appeals categories

In identifying the message types of fear appeals, we first construct a confirmatory topic model from the text of the advertisements. Our first two hypotheses posit the presence of certain fear appeals in the text of IRA sponsored content on Facebook. Accordingly, we needed an approach that precluded the insertion of researcher bias into the organization of those messages. Broadly, there are two competing approaches to quantitative topic modeling: (1) Non-negative matrix factorization (NMF) and the related singular value decomposition (SVD), and (2) latent Dirichlet allocation (LDA). We investigate both options for topic modeling and select an NMF topic model based on our evaluation of a series of fitted models. We found the NMF model to be the most coherent and relevant, producing substantially more orthogonal topic vectors over our input document corpus. NMF is a popular decomposition technique for multivariate data where given a non-negative input matrix, V, two non-negative matrix factors are found, W and H such that their product approximates V with some reconstruction error [60]. NMF begins with an $n \times m$ matrix, V, where m is the number of observations in the data and n is the number of "features." The V matrix is factorized into an $n \times r$ matrix, W, and an $r \times m$ matrix, H, where r is an input parameter specified to be lower than m or n [59]. In this way, NMF can be used to find low rank approximations of the input V matrix.

In its application as a text mining technique, NMF can be used in document summarization or topic modeling when applied over a corpus of text documents. The textual input must be converted into a matrix representation; we elect to use a weighted representation of terms within documents by performing the term frequency dot product inverse document frequency (TF-IDF) transformation over a term frequency matrix. The result is a document × term matrix, with fields indicating the importance of a term to a document given the distribution of that term over all documents in the corpus. The TF-IDF matrix becomes the input, V, to the NMF model.

After removing stop words from the Facebook advertisement text, we generate a term frequency matrix over the documents, limiting the number of terms to 1000. This matrix is subsequently transformed into a TF-IDF matrix prior to non-negative matrix factorization. The number of topics to be extracted, K, is set to 10. The table below illustrates the results of our topic model. For each topic, we list the top 10 weighted terms obtained from the W factor matrix. Using these terms and analyzing the underlying advertisement text, we develop labels for the topics.

Topic model evaluation

The evaluation of topic models is an active area of academic research. Researchers applying topic models have relied on information criterion when using generative probabilistic models such as LDA [61]. However, this approach has been found to produce topic models over text corpora that are not coherent to human judges [62]. We adapt a measure of "topic intrusion" (ibid.), which examines the question of whether human opinion matches the topic mixture vector estimated by a topic model for a given document. We interpret topic intrusion as a coherence measurement that implicitly casts human judgement as the gold standard. More formally, we quantify the agreement between a given topic model and human judgement as follows. First, let denote the vector of topic resonance estimates for the model m given document d. Now, let be an "intruding" topic identified by one of the authors of this paper in the d^{th} document for the m^{th} model. Given the estimated topic resonance for a topic selected by a human judge, and a topic selected by the model, we can compute the total deviation for a given model from human expectation. Using this procedure, we select the 10-topic specification model.

Naturally, the significant additional question pertaining to the topical categories we present below as meaningfully-different types of messaging is whether or not they actually constitute fear appeals, as opposed to less coherent emotional messaging aimed simply at inducing confusion or anger. We qualitatively validate our measurements via assessment of the content of ads in the House data release, which is publicly available. Table 1 above presents a sample of the content we reviewed and our interpretation of the messaging within.

Table 1: Examples of common fear appeals used in IRA Facebook ads.

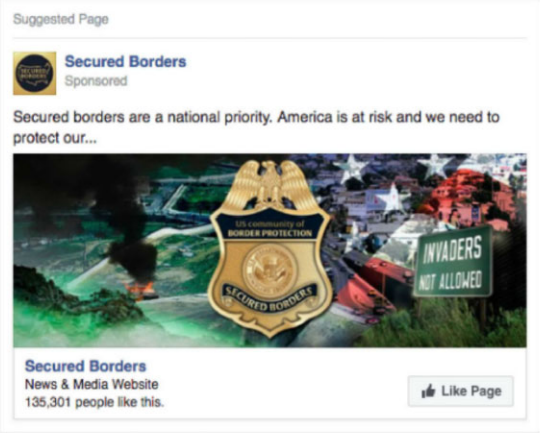
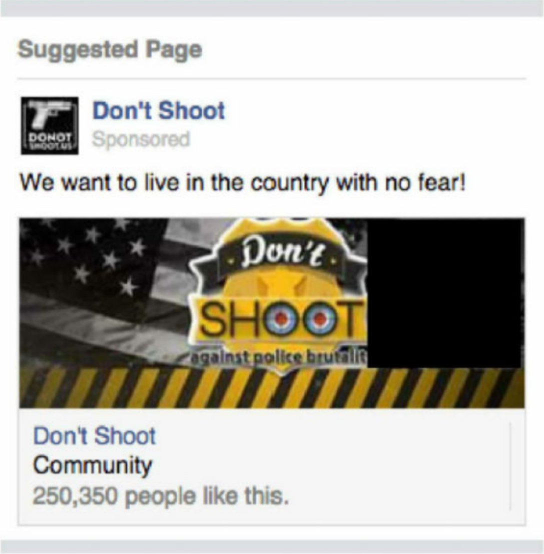
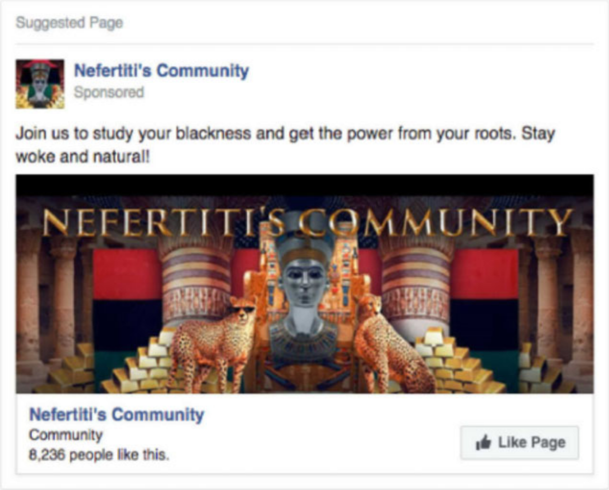
IRA Facebook sponsored content	Fear appeal
 Suggested Page Secured Borders Sponsored Secured borders are a national priority. America is at risk and we need to protect our... US community of BORDER PROTECTION SECURED BORDERS INVADERS NOT ALLOWED Secured Borders News & Media Website 135,301 people like this. Like Page	Invoking a fear of invasion. A *threat severity* appeal.
 Suggested Page Don't Shoot Sponsored We want to live in the country with no fear! [redacted] Don't Shoot Community 250,350 people like this.	Invoking a fear of targeted police violence. A *threat severity* appeal.
 Suggested Page Nefertiti's Community Sponsored Join us to study your blackness and get the power from your roots. Stay woke and natural! Nefertiti's Community Community 8,236 people like this. Like Page	Targeted invocation of a sense of self-efficacy. A *self-efficacy* appeal.

Table 2: Mappings from topics to fear appeals with evidence.

Topic	Fear appeal	Evidence (top terms in topics)
Topic 1	Response efficacy	2nd the community 2nd amendment supporter lover patriot defend gun
Topic 2	Suspected malware	free facemusic stop online music browser player
Topic 3	Threat severity	police brutality officer bm man cop black the video stop police bm stop shoot
Topic 4	Response efficacy	self defense class free feel safe friend event donation basis
Topic 5	Self-efficacy	matter black life join care community blue stand
Topic 6	Self-efficacy	be proud black people want good time say community right join
Topic 7	Self-efficacy	blackexcellence amerikkka africanunity blacknationalism africanandproud panafricanism
Topic 8	Threat severity	like join illegal page dedicate this protection immigrant protection border
Topic 9	Promotion	follow channel instagram facebook subscribe twitter channel channel facebook twitter
Topic 10	Pro-LGBT	member lgbt speak fellow community nation

Table 3: Descriptive statistics.

	Min.	Max.	Mean	Std deviation
Police shooting	1	4	1.382	0.697
BLM protests	0	26	1.874	3.506
Threat severity	0	44	1.464	4.582
Response efficacy	0	5	0.115	0.486
Suspected malware	0	67	0.402	3.882
Self-efficacy	0	44	2.170	4.871
Promotion	0	6	0.179	0.595
LGBT	0	1	0.017	0.129
Clinton–Trump spread	−6.57	13	4.001	3.516
BLM Google trend	0.06	100	4.215	9.830

Clearly, the common phrasing of these advertisements centered on direct questions to the viewer and invitations to take action (i.e. “join us…”) align with the kind of messaging expected within the fear appeals framework.

Topic mappings [60, 61]

We generate a grid of topics, the most important words for each topic, and the most resonant documents for each topic. The authors generated a series of topic parameters and, via reference to several outside assessments, and arrived at those listed in Table 2 as reflecting the most cohesive array of categories.

While we identify self-efficacy, response efficacy, and threat severity messages, it should be noted that these fear appeals are not single-issue appeals. Our labeling of the topics attempts to highlight this fact. For instance, “Topic 10” is labeled “Pro-LGBT” but is actually a self-efficacy appeal. We label it as such so as not to confound effects with those of topics 5–7, which deal with racial injustice-related fear appeals. Finally, in Table 3, we summarize the variables relevant for our analysis.

Empirical model specification

Election proximity effect

It is straightforward to intuit that there is some effect of temporal proximity to the 2016 presidential election on the occurrence of BLM protest activity. Heaney [63], reviewing Gillion [64], points out that elections and protests are closely related; that protest helps consolidate political support, bolster turnout, and accumulate campaign resources. Indeed, as the 2016 presidential election approaches, we notice a large spike in protest activity in our data. However, detecting the point in the series (that is the daily series of protest counts) at which this effect begins to manifest may be challenging. In the study of policy interventions on some aspect of population measured over time, interrupted time-series methodologies have emerged to estimate the causal effects of those interventions. One approach to interrupted time-series analysis is segmented regression. A segmented regression attempts to find breakpoints in the data where relationships between variables of interest may be significantly different. For instance, we may find that for a given segment of the protest time series, Russian IRA advertising emission decisions are significantly different from the rest of the series.

We argue that the relationship between the time to the November 2016 presidential election and the daily count of BLM protests is piece-wise linear. Following Muggeo [65], we fit a segmented regression of the form:

$$g\left(E\left[Y\right]\right) = \alpha Z + \ \beta\left(Z - \varphi\right),$$

where φ represents the (to-be-estimated) breakpoint, β the segmented slope (or the *difference-in-slopes*), Z represents the time to election variable, and α is the slope of the segment of the series for which $Z \leq \varphi$. The term $(Z - \varphi)$ evaluates to zero for all values of Z where $Z \leq \varphi$. Finally, we let $g(E[Y])$ represent the Poisson link function applied to the dependent variable. The above model can easily be extended to include multiple breakpoints. Given this parameterization, the segmented regression can be estimated by fitting a series of linear models to an inputted set of k breakpoints.

Our objective is to identify the point in time when temporal proximity to election day in 2016 may have begun to affect the count of BLM protests. It is *not* to account for non-linearity in the relationship between the time to election and the protests. As such, we prime our segmented regression with starting values of the breakpoints that we visually discern from the chart. We use two starting values for this analysis. The first value is 30 days prior to the November election, and 150 days prior to the November the election. The first value assumes that election proximity effects became pronounced ~30 days prior to the election. The second value assumes that proximity to the party conventions, roughly 120 days before the general elections, may have also interrupted the protest time series. In Fig. 6, these breakpoints correspond to the two major, sustained spikes in BLM protest activity. The exact estimates of the breakpoints in the series are presented in Table 4.

A Davies’ test to reject the null hypothesis that there is no difference in the slopes is significant with P-value 0.002.

Vector auto-regressive models with exogenous variables

Our analysis tracks and seeks to explain the daily advertising emission decisions attributed to the Russian IRA on Facebook using a fear appeals lens. These emission decisions can be represented as a time series. Similarly, our NNMF derived fear appeals categories form daily time series of the volume of ad emissions in each period for each con-

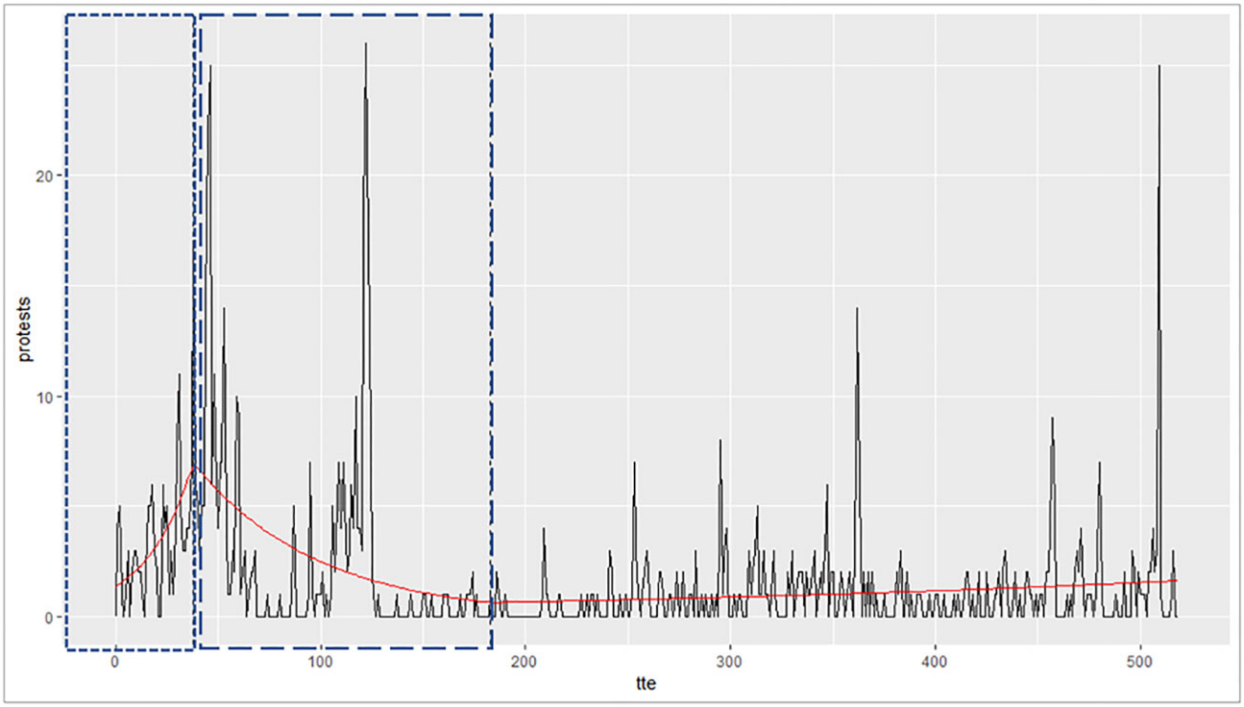


Figure 6: Identified breakpoints and slopes in the protest series.

Table 4: Estimated breakpoints.

Breakpoint parameter	Estimate	Std. error
φ_1	37.989	3.073
φ_2	187.739	9.964

struct. Central to our proposal is that the use of fear appeals by the Russian IRA was systematic, intentional, and time varying. That, insofar as these appeals purported to be driven by the issues (i.e. social justice), they were also driven by a hidden agenda. Accordingly, we model the dynamic interactions and feedback effects of the disinformation campaign by employing a multivariate time-series approach, namely a VARX [66].

The intuition behind this modeling choice rests on the notion that the IRA's decision to emit an advertisement in any one of the categories on a given day is dependent upon their emission decisions in the days before and additionally dependent on their observation of the socio-political conditions in the current period. A VARX model explicitly incorporates *feedback effects*, which capture how levels of the dependent series are affected by lags of itself (e.g. how a decision to emit self-efficacy ads effects future decisions to emit self-efficacy ads), *cross-effects*, which capture the effect of lags of other series in the system on the dependent series (e.g. how a decision to emit self-efficacy ads effects future decisions to emit threat severity ads) and, critically, *contemporaneous effects* of exogenous series on the dependent series (e.g. how fatal police shootings affect the decision to emit self-efficacy ads in the current period) [66]. We may specify the following VARX model, expressed in general form below, following the notational convention in ref. [67]:

$$z_t = \vartheta_0 + \sum_{i=1}^{p} \vartheta_i z_{t-i} + \sum_{j=0}^{s} \beta_j x_{t-j} + \alpha_t,$$

where ϑ_0 is the constant vector, α_t is a sequence of i.i.d disturbances, p and s are positive integers, which represent the lag order of the VARX model (see the section below). The terms z_t and x_t refer, respectively, to k and m dimensional series, where k is the number of endogenous univariate series in the model and m the number of exogenous univariate series in the model. Further, the terms ϑ_i and β_j are $k \times kp$ and $k \times m$ coefficient matrices, respectively. It is important to observe that x_t is free to have contemporaneous effects on z_t. This feature of the VARX extension to standard VAR specifications is critical in our application as we assume that the IRA is able to respond to external stimuli well within a single day.

Lag order selection. An important first step in the estimation of VAR models is the selection of lag-order. In VARX models, this step can be broken up into two stages, where the lag order for the endogenous series is determined first, followed by the lag order for the exogenous series. The lag order of VARX models is typically expressed as $VARX(p, s)$, where s is the lag order of the endogenous series and s the lag order of the exogenous series. The appropriate lags can be determined by evaluating model fit statistics, each estimated based on a different combination of p and s. Generally, we select the most parsimonious model suggested by the information criteria, and in our case, a $VARX(1, 0)$.

Stationarity assumption. A strong assumption in VAR and VARX analysis, one that is critical to consistent identification of the parameters in such models, is that of stationarity. All series in the VARX system should be stationary. A stationary k-dimensional time series is one that exhibits constant covariance and a constant mean. Put another way, the mean and covariance of the series do not depend on time.[7] Several tests exist to examine this assumption. We employ an Augmented Dickey Fuller test for the null hypothesis that a

[7] This condition is actually "weak stationarity" see ref. [66] for details.

Table 5: VAR(1) cross and feedback effects.

z_{t-1}	Threat severity	Response efficacy	Suspected malware	Self-efficacy	Promotion	LGBT
Threat severity	0.005	−0.131	0.032***	−0.0005	−0.015	−2.052**
	(0.011)	(0.110)	(0.006)	(0.011)	(0.072)	(1.005)
Response efficacy	0.074***	0.051	0.016	−0.025	0.150	1.438**
	(0.026)	(0.284)	(0.029)	(0.037)	(0.153)	(0.664)
Suspected malware	−0.127***	0.631***	0.206***	0.176***	−3.256***	−15.32
	(0.036)	(0.196)	(0.015)	(0.026)	(0.353)	(1749)
Self-efficacy	0.039***	−0.163*	0.022***	0.011	0.041	1.088***
	(0.007)	(0.084)	(0.005)	(0.008)	(0.048)	(0.186)
Promotion	0.015	−0.021	0.034**	−0.010	0.062	1.483**
	(0.032)	(0.276)	(0.014)	(0.034)	(0.178)	(0.621)
LGBT	−0.884	−13.58	10.46	−0.149	1.341**	2.622**
	(0.833)	(800.1)	(340.2)	(0.271)	(0.640)	(1.236)

univariate time series is non-stationary and apply the test to all univariate series in our model. For each series, we can reject the null hypothesis with P-values < 0.01.

Vector generalized linear model

We follow the intuition of vector autoregressive models with exogenous variables to model these dynamics. However, our response vector is one of counts and therefore unsuited to standard VARX estimation, which presupposes continuous data in the response and normality in the error terms. As discussed above, we transformed the advertising emissions attributed to the IRA into a multivariate count time series. This transformation began with a thorough consideration of multiple alternative topic models generated over the full collection of Russian IRA attributed Facebook advertisements. Upon identification of the chosen topic model, we analytically mapped the topics to the fear appeals messaging framework. In doing so, we re-casted the raw topic scores in such a way as to discretize each advertising emission. Doing so, results in zt being a multivariate count time series. While the literature has not explored inconsistencies in the estimation of VARX models with multivariate count series, VARX makes explicit normality assumptions that may be violated by our multivariate count response variable. Accordingly, we estimate an alternative specification that modifies relies on generalized linear model.

As our objective is to model the evolving dynamics of the various fear appeals and messaging streams of the Russian disinformation campaign in the 2016 presidential election, we require a modeling framework that enables the estimation of effects associated with those outputs *simultaneously*. Vector generalized linear models enable us to capture these dynamics simultaneously across the various messaging streams while remaining appropriate for our data. GLMs can be described as a special case of VGLMs where there exists only a single linear predictor of the response. We specify an unrestricted (i.e. we do not specify a constraint matrix that forces certain coefficients in the linear predictors to be equal) VGLM that consists of two major components. First, we identify the Poisson distribution as appropriate for our responses and rely on a log link function to relate its expected value to our M linear predictors. Second, we define the M linear predictors, generally, as follows:

$$\eta_j = \sum_{k=1}^{p} \beta_{(j)k} x_k, \;\; j = 1, \ldots, M.$$

We note that $\boldsymbol{x} = (x_1, \ldots, x_p)$ for p explanatory variables. For a given observation in our data, let $\boldsymbol{x}_t$ be a vector of explanatory values for time t for $t, \ldots, T$. We can more descriptively write the linear predictor, η_t, for the t^{th} observation as

$$\eta_t = \begin{pmatrix} \vartheta_{(1)1} \ldots \vartheta_{(1)M} \\ \vdots \\ \vartheta_{(M)1} \ldots \vartheta_{(M)M} \end{pmatrix} z_{t-i} + \begin{pmatrix} \beta_{(1)1} \ldots \beta_{(1)p} \\ \vdots \\ \beta_{(M)1} \ldots \beta_{(M)p} \end{pmatrix} \boldsymbol{x}_t$$

$$+ \begin{pmatrix} \rho_{(1)1} \ldots \rho_{(1)p} \\ \vdots \\ \rho_{(M)1} \ldots \rho_{(M)p} \end{pmatrix} \boldsymbol{x}_{t-1} + \begin{pmatrix} \alpha_{(1)1} \\ \vdots \\ \alpha_{(M)1} \end{pmatrix} elect_t.$$

We thus maintain the AR(1) structure of our VARX model, and allow contemporaneous effects of the explanatory variables $\boldsymbol{x}$. The above model also allows for cross effects and feedback effects (see the coefficient matrix of the lag terms above). Model parameters are obtained with maximum likelihood estimation.

Results

Following maximum likelihood estimation, we summarize the model's coefficients in the tables below. The first table summarizes the cross effects and the feedback effects for those variables intuited to be determined within the system, which is the endogenous variables in VAR terminology. The main diagonal indicates feedback effects, or lagged effects, of one variable on future realizations of itself. Off-diagonal elements show the cross effects, which is the effect of lagged versions of endogenous variables on other endogenous variables. Note that we report standard errors in parentheses. The column variables are the lags. For instance, in Table 5, we can interpret the second element in the first column (0.074) as the lagged effect of an additional threat severity message on response efficacy.

Below, in Table 6, we show the contemporaneous effects of the exogenous variables included in the model. Next, we again examine the effects of the exogenous variables; however, in this case, we estimate lagged effects. Our estimated VGLM suggests that a dynamic understanding of the IRA's use of different fear appeals is revealing. Not only do the findings presented in Tables 5–7 generally bear out our initial expectations regarding the framework of fear appeals for describing the approach taken by sophisticated information warfare operators, but these results also suggest a repetitive lifecycle of audience engagement and messaging centered on the incidence of anchoring events in the targeted country.

We include, in Table 8, the contemporaneous effects of election proximity. Again, this is done to account for variation in the exogenous series that is solely attributable to the politically charged nature of the time periods that are proximal to the election. All fear appeals

Table 6: Contemporaneous effects of exogenous variables.

x_t	Police shooting	BLM Google trend	BLM protests	Clinton–Trump spread
Threat severity	0.403***	0.000	0.082***	−0.049**
	(0.045)	(0.011)	(0.016)	(0.019)
Response efficacy	0.132	−0.057	0.077	0.066
	(0.193)	(0.087)	(0.056)	(0.071)
Suspected malware	0.375**	−0.085	−0.613***	0.037
	(0.155)	(0.119)	(0.144)	(0.061)
Self-efficacy	0.219***	0.018	−0.038**	−0.043***
	(0.039)	(0.011)	(0.019)	(0.017)
Promotion	0.104	0.044	−0.245**	0.036
	(0.148)	(0.062)	(0.098)	(0.056)
LGBT	−0.185	0.056	−0.067	0.345**
	(0.600)	(0.592)	(0.071)	(0.168)

***$P < 0.01$, **$P < 0.05$, *$P < 0.1$.

Table 7: Lagged effects of exogenous variables.

x_{t-1}	Police shooting	BLM Google trend	BLM protests	Clinton–Trump spread
Threat severity	−0.309***	0.022**	−0.168***	0.126***
	(0.061)	(0.010)	(0.028)	(0.020)
Response efficacy	−0.012	0.049	−0.110	0.041
	(0.179)	(0.071)	(0.102)	(0.071)
Suspected malware	0.516***	0.066	0.147*	0.039
	(0.132)	(0.111)	(0.088)	(0.061)
Self-efficacy	−0.096**	0.012	−0.150***	0.094***
	(0.046)	(0.010)	(0.026)	(0.016)
Promotion	0.165	0.008	−0.111	−0.023
	(0.149)	(0.048)	(0.082)	(0.057)
LGBT	−15.27	−0.059	−0.059	−0.117
	(627.5)	(0.498)	(0.498)	(0.153)

***$P < 0.01$, **$P < 0.05$, *$P < 0.1$.

Table 8: Contemporaneous effects of election proximity.

$elect_t$	Threat severity	Response efficacy	Suspected malware	Self-efficacy	Promotion	LGBT
Election dummy	1.602***	0.883**	2.301***	0.941***	1.692***	−13.55
	(0.113)	(0.436)	(0.325)	(0.103)	(0.353)	(1396)

related message emissions are significantly ramped up in periods that are proximal to the election.

Finally, we conduct non-linear granger causality tests and report the results below. Our non-linear tests rely on a multi-layer perceptron neural network (MLP) that is insensitive to the count structure of our data [68]. This procedure first fits a univariate predictive multilayer perceptron neural network and subsequently fits a bivariate MLP. If the bivariate model provides better forecasts than the univariate model (with respect to the series in the univariate model), we can reject the test's null hypothesis of non-granger causality. As our ADF stationarity tests hold, we do not perform differencing prior to conducting these tests. Consistent with the above models, we use a single period lag. We test each pairwise relationship in our model for both endogenous and exogenous series. In Table 9, we only present significant or marginally significant results for conciseness. These results are with respect to a single period lag.

Table 9: Significant and marginal results from tests for Granger causality.

From	To	*F*-statistic	*P*-value
Clinton spread	*Self-efficacy*	4.121	0.043**
Threat severity	*Response efficacy*	10.914	0.001***
Response efficacy	*Suspected malware*	4.011	0.046**
Self-efficacy	*Response efficacy*	3.344	0.068*
Promotion	*Response efficacy*	3.231	0.073*
LGBT	*Response efficacy*	3.136	0.077*

With regards to the sequence of IRA messaging, the results in Table 5 show that threat severity messaging is positively and significantly associated with efficacy messaging during the proceeding period (H_1). Importantly, this is not an in-kind relationship; threat severity messaging does not appear to be linked to efficacy messaging in the preceding period, suggesting that IRA operators made—and, given the entity's ongoing operations, still likely make—the conscious decision to emphasize threat mitigation messaging in response to initial audience-facing stimuli. Further, we find that this relationship is granger causal especially with respect to H_{1a}. Moreover, comparative analysis of Tables 6 and 7 suggests that IRA operators consciously crafted messaging around domestic triggering events (H_{2a} and H_{2b}). While Table 6 shows a strong, positive, contemporaneous association between threat messaging and incidences of police shootings and BLM protests, the lagged result in Table 7 shows a negative and significant relationship 1 day later. Accordingly, we find partial support for H_{2a} and $H_{2b.}$ That is, while we hypothesized that the increasing severity of the focal macro-social issues would lead to increasing emissions of both threat and efficacy appeals, we find that of the three macro-social issue variables in the model, only the contemporaneous effect of an increase in police shootings has positive coefficients for threat severity and self-efficacy. This makes considerable sense. As both Congressional and journalistic findings in investigations into Russian activity on social media have shown, Facebook ad purchases and publication almost always occurred within an extremely short span of time. Ad buys that often referenced real world developments from just the previous 24 hours would go through mere hours before content was presented to users. In other words, the lifecycle of IRA content generation and distribution via Facebook was a daily one. Thus, here, it is logical that we see strong threat messaging in periods where domestic crisis events take place only to disappear in the following period as IRA operators pivot from the task of threat inflation to spin real world circumstance.

With respect to H_3, we again find some support in the estimated model. There is a positive relationship between unfavorable states of the hidden agenda on the emission of threat severity messages (Table 7). Recalling that the ICA [69] assessed that the hidden objective of this coordinated trolling activity was the defeat of Hillary Clinton in the 2016 presidential elections, a unit increase in the Clinton–Trump spread is an unfavorable state. However, our expectation only holds for the lagged effect as the contemporaneous effect is significant, small, and negative (Table 6). Similarly, we find some support for H_4. Our results indicate that the lagged effect of unfavorable states in the hidden agenda are not only positive but also significant, with respect to efficacy appeals. The direction of this effect is confirmed in Table 9 (see the first row) using tests for Granger causality. Again, the effect is reversed in the contemporaneous case (although this reversal is not Granger-causal). One possible explanation for these findings is that the troll operators did not react instantaneously to poll results. We note that both threat severity and self-efficacy effects are negative, significant, and close to zero in the contemporaneous series.

Functional efforts by the IRA to "capture" audiences and promote their messaging also follow the broad pattern we outline in our expectations above, though in at least one unexpected fashion. The results above suggest that attempts to expand the viewership of IRA content took at least two formats. Broadly, Facebook ad content promoted IRA accounts and pages with structured content asking viewers to (1) visit webpages or (2) follow accounts on Twitter, Instagram, and elsewhere. Of particular interest, this attempt to engender a larger following for IRA accounts and the causes being represented in messaging is not linked to the pillars of the IRA fear appeal effort. The "Promotion" topic is not positively and significantly linked with other themes with a single exception—incidence of LGBT content in ads. Analysis of the LGBT topic results thus tell us about the IRA's general attempts to promote their accounts. Simply put, the appearance of LGBT content is clearly *not* linked with directed efforts to capitalize on audience fears. Rather, this content appears in ad content only in periods after efficacy messaging has appeared or where a greater Clinton–Trump spread in the poll is evident. This suggests that such content was the yin to the yang of more severe threat severity substance in periods where no triggering event existed to necessitate efficacy messaging. As such, it seems logical that generalized promotion efforts were targeted only to periods where no attempt to inflate and inflame audience reactions via fear appeals was underway. This supports our assumption captured in H_2.

And yet, we also see efforts to capitalize on fear appeals to promote the cause. The logic of our initial assumption, again, was that such messaging would be followed by short-term assertive attempts to expand influence off the basis that anybody paying attention to IRA messaging during the secondary stages of the fear communication lifecycle would be particularly prone to suggestion therefrom. This follows the general logic of phishing campaigns [70] where the purpose of mass-produced communication is generally to identify gullible or otherwise susceptible populations for further exploitation.

Here, we see such assertive attempts at expanded influence in the form of malware pushes enabled by Facebook ad content. Our topic model captured efforts to promote several applications—one in particular in the form of FaceMusic—that have since been identified as malicious in nature. Specifically, these applications contained code linked directly to known click fraud crime where the point is to leverage a victim computer to clandestinely upvote content on one or numerous platforms. This approach to click manipulation is nigh undetectable to any web administrator (as it appears that the popularity of a piece of content stems from the actions of multiple users with unique IP addresses) and is useful for promoting otherwise fabricated content. Here, two results bear particular attention. First, malware-pushing content is negatively and significantly linked to threat severity messaging but positively and significantly linked to efficacy messaging. This suggests that the effort to spread influence-enhancing malware was consciously targeted to the operational juncture where operators perceived the most opportune audience composition for proliferation. Second, malware-pushing content has a strong negative link to more general promotion, strengthening our suggestion that such efforts were consciously decoupled from the inflammatory fear appeal campaign. Figure 7, above, illustrates this decoupling as a lifecycle model of attack practiced by IRA operators.

Discussion and implications: a lifecycle of IO social media engagement

We have argued that the outsider-looking-in dynamic of state-sponsored IO and CEIO means that belligerents necessarily rely on domestic events to provide the context and initial interest in current events from which influence can be built, expanded, and leveraged to create tangible effects. To validate and illustrate this framework, we examine the case of Russian active measures undertaken by the IRA in the USA from 2015 to 16. Our results provide evidence in support of our hypotheses, though in some cases support comes in unexpected forms. We see clear evidence that macrosocial triggering events prompt the onset of behavioral changes on the part of the IRA in this case. Moreover, we find that the kind of messaging evolves around such events. That said, several results suggest that yet further work needs to be done better establish and flesh out this logic of influence from afar. Following both police shootings and incidence of BLM events, for instance, efficacy messaging was less likely to appear in Facebook ad content than expected, though these results are either statistically insignificant or relatively weak.

We suggest that such outcomes are to be expected given the complexities of running a broad-scoped influence operation from

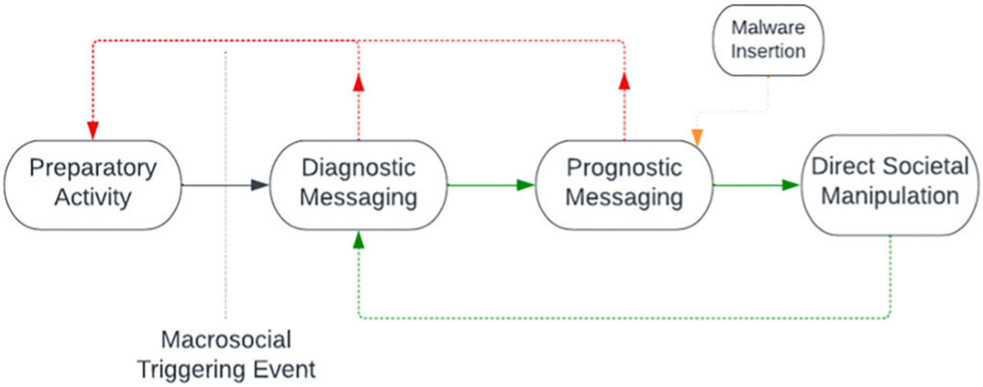


Figure 7: A lifecycle of IO social media engagement based on domestic conditions.

overseas and do not, in any case, fundamentally contradict the model of tactical approach being outlined in our work here. That model is reasonably straightforward—a lifecycle of audience outreach and agenda-setting that attempts to employ threat messaging and efficacy selling to inflame and divide American discourse following a necessary triggering event. To this last point, the attack lifecycle we suggest and see in the data are not a continuous or repetitive one as might be true for a legitimate domestic political actor (i.e. for IO, some rhetorical sequence of messages does not obviously drive the next). Again, this dynamic simply follows the reality of influence campaigns as clandestine and based beyond American territory. The IRA clearly relied on triggering events to enact their strategy, employing at other times simple attempts to generically shape discourse on prescient issues, offering competing voices to a divided population, and generally promoting their own accounts.

Coordinated state-sponsored IO that aim to antagonistically disrupt the organic flow of information in societies are clearly of substantial, growing concern to citizens of the world's democracies. In our study, we have demonstrated that the generation of message content by IO operatives can be understood by the state of their hidden agenda at a given time and is mediated by the state of the macro-social division(s) that they seek to exploit. Methodologically, we show that fear appeals are a viable lens for characterizing the inauthenticity of what many simply think of as coordinated trolling behavior and have, thus, fired an opening salvo for future research into the automated, early detection of these campaigns. More broadly, there are significant implications of our model of policymakers and scholars that seek to neutralize the impact of political interference. The lifecycle suggested by our model clearly suggests, as much research on global cyber conflict increasingly has done, that observation and consideration of sociopolitical context is critical for the effective operationalization of efforts to predict and mitigate foreign-based IO threats. We also make a major contribution, albeit unexpected from the outset, to the evolving research program on cyber-enabled IO (or CEIO) insofar as our empirical testing illustrated how the IRA embedded an effort to spread influence-building malware into their messaging lifecycle. Cyber operations and influence campaigns are clearly not just the close relatives that many scholars have described, but have real and prospectively impactful operational correlates. Finally, there are clear implications for the management and design of the social media systems emphasized in so much modern IO. Specifically, the transition from influence generation to influence in action is clearly a more critical factor explaining where some IO might succeed in causing tangible effects than is the simple fact of presence on a platform like Twitter or Facebook. Going forward, both policymakers and technology developers would do well to focus on the kinds of speech that are closely linked to this attempt to transform voice to, among other things, violence.

Funding

The authors have no funding to report.

Author contributions

Ugochukwu Etudo (Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Software [equal], Visualization [equal], Writing – original draft [equal], Writing – review & editing [equal]), Christopher Whyte (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Project administration [equal], Supervision [equal], Writing – original draft [equal], Writing – review & editing [equal]), Victoria Yoon (Project administration [equal]), and Niam Yaraghi (Project administration [equal])

Conflict of interest statement. The authors have no conflicts of interest to declare. All co-authors have seen and agree with the contents of the manuscript.

References

1. Bennett WL, Livingston S. The disinformation order: disruptive communication and the decline of democratic institutions. *Eur J Commun* 2018;**33**:122–39.
2. Howard PN, Ganesh B, Liotsiou D. *TheIRA, Social Media and Political Polarization in the United States, 2012–2018*. 2018. https://research.rug.nl/en/publications/the-ira-social-media-and-political-polarization-in-the-united-sta.
3. Golovchenko Y, Buntain C, Eady G,. *et al.* Cross-platform state propaganda: Russian trolls on twitter and YouTube during the 2016 US Presidential Election. *Int J Press/Politics* 2020;**25**(3):357–89.
4. Fang Y, Gao J, Huang C,. *et al.* Self multi-head attention-based convolutional neural networks for fake news detection. *PLoS One* 2019;**14**: 1–13.
5. Rid T. *Active Measures: The Secret History of Disinformation and Political Warfare*. New York, NY: Farrar, Straus and Giroux, 2020.
6. Vicic J, Gartzke E. "Cyber-enabled influence operations as a 'center of gravity' in cyber conflict: the example of Russian foreign interference in the 2016 U.S. Federal election." 2022. Working paper.
7. Papanastasiou Y. Fake news propagation and detection: A sequential model. *Manage Sci* 2020;**66**:(5)1826–46.
8. Kim A, Dennis AR. Says who? The effects of presentation format and source rating on fake news in social media. *MIS Q Manag Inf Syst* 2019;**43**:1025–39.
9. Pennycook G, Rand DG. The psychology of fake news. *Trends Cogn Sci* 2021;**25**:388–402.

10. Di Domenico G, Sit J, Ishizaka A,. *et al.* Fake news, social media and marketing: a systematic review. *J Bus Res* 2021;**124**:329–41.
11. Kim A, Moravec PL, Dennis AR. Combating fake news on social media with source ratings: the effects of user and expert reputation ratings. *J Manag Inf Syst* 2019;**36**:931–68.
12. Papanastasiou Y, Savva N. Dynamic pricing in the presence of social learning and strategic consumers. *Manage Sci* 2017;**63**:919–39.
13. Witte K. Putting the fear back into fear appeals: the extended parallel process model. *Commun Monogr* 1992;**59**:329–49.
14. Zannettou S, Sirivianos M, Blackburn J,. *et al.* The web of false information: rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J Data Inf Qual* 2019;**11**:10.
15. Zannettou S, Caulfield T, De Cristofaro E,. *et al.* Disinformation warfare: understanding state-sponsored trolls on Twitter and their influence on the web. *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*. 2018.
16. Möldera H, Sazonovb V. Information warfare as the hobbesian concept of modern times—the principles, techniques, and tools of Russian information operations in the donbass. *J Slav Mil Stud* 2018;**31**:308–28.
17. Kaplan S. Framing contests: strategy making under uncertainty. *Organ Sci* 2008;**19**:729–52.
18. Stewart LG, Arif A, Starbird K. Examining trolls and polarization with a retweet network. *Proc WSDM Work Misinformation Misbehavior Min Web (MIS2)* 2018:**6**. http://faculty.washington.edu/kstarbi/examining-trolls-polarization.pdf.
19. Levin I, Sinclair JA, Alvarez RM. Participation in the wake of adversity: blame attribution and policy-oriented evaluations. *Polit Behav* 2016;**38**:203–28.
20. Levin DH. *Meddling in the Ballot Box: the Causes and Effects of Partisan Electoral Interventions*. New York, NY: Oxford University Press, 2020.
21. Whyte C. Cyber conflict or democracy "hacked"? How cyber operations enhance information warfare. *J Cybersecurity* 2020;**6**:tyaa013.
22. Chivvis CS. Understanding Russian "Hybrid Warfare". *Rand Corp* 2017;**17**.
23. Whyte C, Thrall AT, Mazanec BM. *Information Warfare in the Age of Cyber Conflict*. Abingdon, Oxfordshire, United Kingdom: Routledge, 2020.
24. Foote C, Maness RC, Jensen B,. *et al.* Cyber conflict AT the intersection of information operations. In: *Information Warfare in the Age of Cyber Conflict*. Abingdon, Oxfordshire, United Kingdom: Routledge, 2020, 54–69.
25. Libicki MC. The convergence of information warfare. In: *Information Warfare in the Age of Cyber Conflict*. Abingdon, Oxfordshire, United Kingdom: Routledge, 2020, 15–26.
26. Whyte C. Beyond tit-for-tat in cyberspace: political warfare and lateral sources of escalation online. *Eur J Int Secur* 2020;**5**:195–214.
27. Way LA, Casey A. Russia has been meddling in foreign elections for decades. Has it made a difference? *The Washington Post*. 2018. https://www.washingtonpost.com/news/monkey-cage/wp/2018/01/05/russia-has-been-meddling-in-foreign-elections-for-decades-has-it-made-a-difference/ (23 March 2021, date last accessed).
28. O'Connor C. New technologies and the law in war and peace. *Glob Change Peace Secur* 2020;**32**:234–5.
29. Keller T, Graham T, Angus D,. *et al.* 'Coordinated inauthentic behaviour' and other online influence operations in social media spaces. *AoIR Selected Papers of Internet Research*. 2020.
30. Zannettou S, Caulfield T, Cristofaro E De. Disinformation warfare : understanding state-sponsored trolls on Twitter and their influence on the web. *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*. 2017.
31. Song J, Fiore SM. For whom the tale's told: towards a multidimensional model of targeted narrative persuasion in information operations. *Proc Hum Factors Ergon Soc Annu Meet* 2020;**64**:1505–9.
32. Linvill DL, Boatwright BC, Grant WJ,. *et al.*"THE RUSSIANS ARE HACKING MY BRAIN!" investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign. *Comput Human Behav* 2019;**99**:292–300.
33. Dawson A, Innes M. How Russia's internet research agency built its disinformation campaign. *Polit Q* 2019;**90**:245–56.
34. Lukito J, Suk J, Zhang Y,. *et al.* The wolves in sheep's clothing: how Russia's Internet Research Agency tweets appeared in US news as vox populi. *Int J Press* 2020;**25**:196–216.
35. Fisher A. A new Cold War? International public opinion of Russia and the United States. *Int J Public Opin Res* 2020;**32**:143–52.
36. Brantly AF. A brief history of fake: surveying Russian disinformation from the Russian Empire through the Cold War and to the present. *Information Warfare in the Age of Cyber Conflict*. Abingdon, Oxfordshire, United Kingdom: Routledge, 2020, 27–41.
37. Brantly A. Battling the bear. In: *Cyber Security Politics*, **157**. Abingdon, Oxfordshire, United Kingdom: Routledge, 2022.
38. Kerr J. The Russian model of digital control and its significance. *Artificial Intelligence, China, Russia, and the Global Order Air University Press*. 2018, 55.
39. Lin H, Kerr J. On cyber-enabled information/influence warfare and manipulation. *Center for International Security and Cooperation (CISAC)*. 2017.
40. Vargo CJ, Hopp T. Fear, anger, and political advertisement engagement: a computational case study of Russian-linked Facebook and Instagram content. *Journal Mass Commun Q* 2020;**97**:743–61.
41. Spangher A, Ranade G, Nushi B,. *et al.* Analysis of strategy and spread of Russia-sponsored content in the US in 2017. *arXiv:1810.10033*. 2018.
42. Silva M, Giovanini L, Fernandes J,. *et al.* Facebook ad engagement in the russian active measures campaign of 2016. *arXiv:2012.11690*. 2020.
43. Tannenbaum MB, Hepler J, Zimmerman RS,. *et al.* Appealing to fear: a meta-analysis of fear appeal effectiveness and theories. *Psychol Bull* 2015;**141**:1178.
44. Jamieson KH. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. New York, NY: Oxford University Press, 2018.
45. Klausen J. Tweeting the Jihad: social media networks of western foreign fighters in Syria and Iraq. *Stud Confl Terror* 2015;**38**:1–22.
46. Janis IL. Effects of fear arousal on attitude change: recent developments in theory and experimental research. *Adv Exp Soc Psychol* 1967;**3**:166–224.
47. Leventhal H. Findings and theory in the study of fear communications. *Adv Exp Soc Psychol* 1970;**5**:119–86.
48. Rogers RW. A protection motivation theory of fear appeals and attitude change1. *J Psychol* 1975;**91**:93–114.
49. Rogers RW, Deckner CW. Effects of fear appeals and physiological arousal upon emotion, attitudes, and cigarette smoking. *J Pers Soc Psychol* 1975;**32**:222.
50. Lazarsfeld PF, Berelson B, Gaudet H. The people's choice: how the voter makes up his mind in a presidential campaign. Columbia University Press, 1944.
51. Katz E, Lazarsfeld PF, Roper E. *Personal influence: The part played by people in the flow of mass communications*. Abingdon, Oxfordshire, UK: Routledge, 2017.
52. Maloney EK, Lapinski MK, Witte K. Fear appeals and persuasion: a review and update of the extended parallel process model. *Soc Personal Psychol Compass* 2011;**5**:206–19.
53. Snow DE, Benford R. Ideology, frame resonance, and participant mobilization. *Int Soc Mov Res* 1988;**1**:197–218.
54. Pickering A. The mangle of practice : agency and emergence in the sociology of Science1. *Am J Sociol* 1993;**99**:559–89.
55. de Hoog N, Stroebe W, de Wit JBF. The impact of vulnerability to and severity of a health risk on processing and acceptance of fear-arousing communications: a meta-analysis. *Rev Gen Psychol* 2007;**11**:258–85.
56. Lianos M. *Dangerous Others, Insecure Societies: Fear and Social Division*. Abingdon, Oxfordshire, UK: Routledge, 2016.
57. Boehnke K, Macpherson MJ, Meador M,. *et al.* How West German adolescents experience the nuclear threat. *Polit Psychol* 1989;**10**:419–43.
58. Boehnke K, Kindervater A, Baier D,. *et al.* Social change as a source of macrosocial stress: does it enhance nationalistic attitudes? A

cross-cultural study of effects of the EU eastern enlargement. *Eur Soc* 2007;**9**:65–90.

59. Farkas J, Bastos M. State propaganda in the age of social media: examining strategies of the Internet Research Agency. *7th European Communication Conference (ECC), Lugano, Switzerland (31 October–3 November)*. 2018.
60. Lee D, Seung H. Algorithms for non-negative matrix factorization. *Adv Neural Inf Process Syst* 2001:556–62.
61. Hwang EH, Singh PV, Argote L,. *et al*. Jack of all, master of some: information network and Innovation in crowdsourcing communities. *Inf Syst Res* 2019;**30**:389–410.
62. Chang J, Boyd-Graber J, Gerrish S,. *et al*. Reading tea leaves: how humans interpret topic models. *NIPS'09: Proceedings of the 22nd International Conference on Neural Information Processing Systems*. 2009, 1–9.
63. Heaney MT. Review essay: connecting elections and protests. *Interest Groups Advocacy* 2020;**9**:552–5.
64. Gillion DQ. *The Loud Minority: Why Protests Matter in American Democracy*. Princeton, NJ: Princeton University Press, 2020.
65. Muggeo VRM. Segmented: an R package to fit regression models with broken-line relationships. *R News* 2008;**3**:343–4.
66. Luo X, Zhang J, Duan W. Social media and firm equity value. *Inf Syst Res* 2013;**24**:146–63.
67. Tsay RS. *Multivariate Time Series Analysis: With R and Financial Applications*. Hoboken, NJ: John Wiley & Sons, 2013.
68. Tank A, Covert I, Foti N,. *et al*. Neural granger causality for nonlinear time series. *IEEE Trans Pattern Anal Mach Intell* 2018;**44**: 4267–79.
69. National Intelligence Council. *Assessing Russian Activities and Intentions in Recent US Elections*. 2017, 14.
70. Garera S, Provos N, Chew M,. *et al*. A framework for detection and measurement of phishing attacks. *Proceedings of the 2007 ACM Workshop on Recurring Malcode*. 2007, 1–8.