Foreign Policy at BROOKINGS

THE BREAKOUT SCALE: MEASURING THE IMPACT OF INFLUENCE OPERATIONS

BEN NIMMO

SEPTEMBER 2020

EXECUTIVE SUMMARY

One of the greatest challenges in the study of disinformation and influence operations (IOs) is measuring their impact. IO practitioners acknowledge that measuring the impact of their own operations is a complex process that requires careful study and calibration; it is much harder for operational researchers, whose job it is to identify and expose IO, without reliable information on what the operation is trying to achieve.

This paper seeks to answer that challenge by proposing "The Breakout Scale," a comparative model for measuring IOs based on data that are observable, replicable, verifiable, and available from the moment they were posted. It is intended for use by the operational research community for real-time categorization of IOs as they are identified.

The breakout scale divides IOs into six categories, based on whether they remain on one platform or travel across multiple platforms (including traditional media and policy debates), and whether they remain in one community or spread through many communities. At the lowest end of the spectrum, Category One operations only spread within one community on one platform, while Category Two operations either spread in one community across multiple platforms, or spread across multiple communities on one platform. Category Three operations spread across multiple social media platforms and reach multiple communities.

Category Four operations break out from social media completely and are amplified by mainstream media, while Category Five operations are amplified by high-profile individuals such as celebrities and political candidates. An IO reaches Category Six if it triggers a policy response or some other form of concrete action, or if it includes a call for violence.

The scale is designed to allow operational researchers to compare the probable impact of different operations in real time and on the basis of measurable and replicable evidence. It also underscores the importance of mainstream journalists, policymakers, and celebrities. Such high-profile influencers can play a pivotal role in bringing IOs to new audiences: It will be important to raise their awareness of the ways in which they can themselves be targeted by influence operators.

THE INFLUENCE OPERATION CHALLENGE

One of the greatest challenges in the study of disinformation and influence operations (IO) is measuring their impact.[1] The purpose of this paper is to propose a comparative model for operational researchers — those who seek to study and expose live IOs in real time, whether open-source analysts, journalists, or in-platform investigators — to calibrate their assessments of an operation's impact, and to compare the impact of different operations according to a common analytical framework that goes beyond the raw numbers of social media engagement.

IO practitioners acknowledge that measuring the impact of their own operations is a complex process that requires careful study and calibration. The U.S. Army manual on Information Operations devotes nine pages to impact assessment, with an emphasis on measurable changes in behavior.[2] Similarly, the Internet Research Agency (IRA) in St. Petersburg was obsessed with impact metrics, both on the production side (number of posts and comments per day) and on the consumption side (numbers of likes and retweets), according to former Russian trolls.[3]

If it is difficult for the people running an operation to define how well they are performing, it is even more difficult for outside observers. First, they do not necessarily know what that operation was trying to achieve. As an example, on May 21, 2016, the IRA covertly organized two competing rallies in Houston, Texas. One side was protesting an Islamic cultural center, the other was counter-protesting. According to a Houston journalist, the Facebook invitation to the "Stop Islamization of Texas" protest ended: "Feel free to bring along your firearms, concealed or not!"[4]

At first sight, it is a shocking story, and it captured analysts' attention when it was exposed in November 2017. The initial assumption, including by this author, was that the troll operation was a success: Working from some 5,500 miles away, it managed to mobilize two adversarial, potentially armed, groups in America. However, with hindsight, that assumption was questionable at best. Turnout for the rallies was small: There were under a dozen anti-Muslim protesters, few of them armed, and only an estimated 60 counter-protesters.[5] Nobody was hurt, and the only arrest reported was of a woman who failed to get out of the way of a maneuvering police car.[6] Was the turnout a success for the operators, or was the lack of violence a disappointment? Without knowing what the Russian trolls wanted to achieve, it is impossible to know whether they achieved it.

A second challenge is that IOs are typically aimed at influencing the target audience's sentiment. The traditional way of measuring that influence would be through repetitive opinion polling before, during, and after the campaign.[7] Sentiment analysis of comments on social media can provide a proxy for polling,[8] but is, by definition, limited to those who actually commented, and cannot provide information on those who saw a piece of content but did not react publicly. Large-scale sentiment analysis is further complicated by the different conversation structures and data access limitations of different platforms.[9] Thus, even if researchers can identify the sentiment that the IO sought to influence, it is difficult to observe and measure changes in that sentiment across the target audience.

Third, geopolitical IOs tend to be aimed at influencing sentiment on large-scale, strategic subjects. State-sponsored IOs have, for example, sought to depress voter support for Hillary Clinton,[10] discredit the Hong Kong protesters[11] and the Syrian Civil Defence (also known as the White Helmets rescue group),[12] increase support for the Iranian state[13] and the Honduras government,[14] discredit critics of hard-line Western policies towards Iran,[15] and promote pro-Saudi, anti-Iranian sentiment.[16] These are complex topics where many other factors also impact audience sentiment, and the chains of cause and effect are tangled. Operational researchers are

unlikely to be able to disambiguate the different factors and identify the exact role played by IOs while the influence attempt is underway.

Finally, investigators very seldom come across IOs right at the start. Far more commonly, they encounter the IO in progress and have to devote considerable time to working out when the operation started and what assets were being deployed. This means that the investigators typically work off partial and evolving datasets, which inevitably affects definitive impact measurement.

APPROXIMATING IMPACT

Researchers have developed a number of solutions to these challenges. One common approach is to focus on the raw metrics of social media engagement, such as the number of likes or shares a specific piece of content received. This approach treats reach as a proxy for impact, and it can give an impression of scale. For example, several news outlets reported that as many as 126 million American Facebook users could have seen IRA-created content on that platform between January 2015 and August 2017.[17] (Some researchers attempt to nuance these figures by comparing the IO they are exposing with the reach of genuine news outlets, but this runs the risk of the researcher choosing an inappropriate comparison which makes "their" IO look more important.)

However, traffic numbers on a single platform are a poor approximation for an IO's overall impact, especially when the IO works across many platforms: In the case of the IRA's 126 million Facebook views, for example, the figure took no account of IRA content on Twitter, Instagram, YouTube, and various other websites. Aggregating the numbers across different platforms without providing further context can also prove misleading: The IRA posted over 9 million tweets between 2014 and 2017,[18] but more of these were in Russian than in English, making the overall number a poor indicator of impact. The numbers also fail to factor in any artificial amplification by fake accounts, the size of the audience segments that were targeted, and the number of times that each user may have been exposed to the operation's content. Finally, the numbers give no indication of whether offline citizens — those who are not on social media — were also exposed to IRA messaging. The aggregate number of 126 million potential viewers confirms that the IRA was running a big operation, but it cannot be used to measure meaningful impact.

Engagement numbers are not the only tool available. Researchers can, for example, use internet search results to illustrate how much of the conversation is dominated by the IO,[19] or examine the spread of specific artifacts created by the IO, such as memes, videos, hashtags, or unique word formations.[20] These approaches provide crucial insights into the performance of individual operations or moments within those operations, but they lack a standardized scale of measurement which would allow researchers to compare different operations.

This paper seeks to fill that gap by proposing "The Breakout Scale," a comparative model for measuring IO based on data that are observable, replicable, verifiable, and available from the moment they were posted. The operational research community can use this scale for real-time categorization of IO as they are identified.

PRINCIPLES OF THE BREAKOUT SCALE

This paper takes the view that, for researchers into live operations, it is not practically possible to measure sentiment change, and thus to arrive at a direct assessment of impact. However, if an influence operation is to change the beliefs or behavior of a community, it has to be able to land its content in front of that community first, and the way a message passes from one community to another *can* be tracked and measured. The Breakout Scale seeks to define how effectively the content launched by influence operations spreads in three dimensions — on social media, in the mainstream media, and in real life — using factors that can be compared across different operations.

Following the concept proposed by IO researcher Alicia Wanless, the scale considers the information environment as a complex ecosystem[21] in which each social media platform is a discrete entity that exists alongside the others, and alongside many other sources of information and communication, such as radio, TV, print newspapers, blogs, political declarations, and legislative initiatives.

The most dangerous influence operations will be those that show the greatest ability to spread to many different communities, across many platforms, and into real-life discourse.

A story planted by a disinformation actor can be considered akin to a virus that is inserted into the ecosystem.[22] As in nature, the virus will typically start by attempting to infect one entity. However, even if it fully infects that entity, the entity in question only represents a small percentage of the total ecosystem. For a virus to spread throughout the ecosystem, it must be able to jump from one entity to another.[23] The most potentially dangerous viruses will thus be those that show the greatest ability to jump from one entity to another. Or, to return to our subject, the most dangerous influence operations will be those that show the greatest ability to spread to many different communities, across many platforms, and into real-life discourse.

The key steps in assessing a particular operation are therefore to identify *where* the IO actor first planted it (the "insertion point(s)")[24] and *whether* it managed to break out of that particular ecological niche to infect other entities (the "breakout moment(s)").

The insertion point can be a community on a single platform. This was the case in Saudi Arabia's Twitter operations against Qatar in the summer of 2017, during which Saudi royal adviser Saud al-Qahtani repeatedly tweeted anti-Qatar hashtags: These scored high traffic numbers, but the shares were primarily among existing Saudi supporters, to judge by their profiles and the content they shared.[25] It can be a broader community that is active across multiple platforms. This was the case with the Black Lives Matter movement, which the IRA targeted across Facebook, Twitter, Instagram, and blog posts from 2015 onwards.[26] It can be an individual whom the IO approaches with personally tailored messages, emails, or mentions on social media, as in the case of the Iranian operation "Distinguished Impersonator," which used fabricated journalist personas to approach real individuals for interviews.[27] One operation can attempt different insertion points at different stages of the process, such as the Russian operation "Secondary Infektion"[28] that leaked U.K.-U.S. trade documents ahead of the British election in 2019: This started on Reddit, moved to tagging politicians and journalists on Twitter, and ended by emailing political activists directly.[29]

Breakout moments occur when an influence operation's message spreads organically from the insertion point into new communities.[30] This represents an escalation of the operation's potential impact, as it introduces it to parts of the information ecosystem where it was not present before.

Breakout moments can be *on-platform*, when a user at the insertion point picks up a message from the IO and repeats it to a new audience (e.g., a genuine Black Lives Matter activist retweeting an IRA meme). They can be *cross-platform*, when a user at the insertion point posts a message from the IO onto a platform where that IO has no presence (e.g., an American Confederate user taking a screenshot of an IRA tweet and posting it on Instagram, or a pro-China user taking a clip from a pro-government video on YouTube and sharing it on TikTok). They can be *cross-medium*, when messages or content that started on social media are picked up and reproduced by traditional media[31] (such as the many news outlets around the world that embedded IRA tweets in their reports).[32] They can land their messages with *major influencers*, when a high-profile figure unaffiliated with the operation (such as a politician or celebrity) repeats the operation's

message, especially if they explicitly endorse it. At the very top of the scale, they can have a *policy impact*, if decisionmakers shape or change their policies as a result of the IO's work, or carry a *call to violence*, which gives them an imminence and urgency they might not otherwise have.[33]

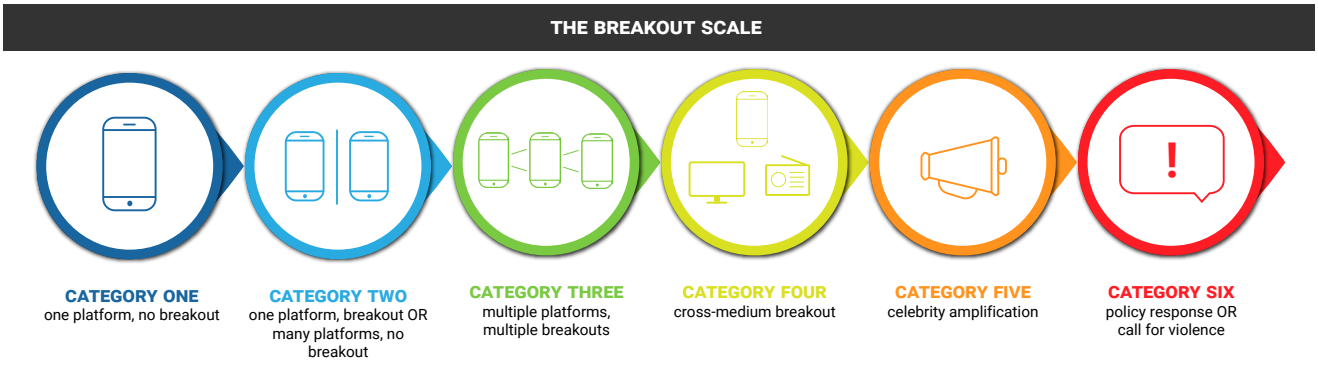
One operation can have successive breakout moments. For example, in 2016, Russia's military intelligence service, commonly known as the GRU, hacked into the servers of the Democratic National Committee (DNC) and downloaded thousands of emails.[34] The operation that followed first created a false "hacktivist" persona (insertion point) which attracted followers (on-platform breakout followed by cross-platform breakout), and reached out to individual journalists to pitch the leaks to them (cross-medium breakout). WikiLeaks then contacted the Russian operation and offered to host future leaks on its behalf, arguing that "it will have a much higher impact than what you are doing" (breakout via a major influencer).[35] Ultimately, the release of the DNC emails led to the resignation of DNC Chair Debbie Wasserman Schultz and other senior DNC staff (breakout manifesting in policy impact).[36]

Each successive breakout moment brings the IO's content to a larger and/or more influential audience. As such, these breakout moments form a rising scale of potential impact, from one user community, to many online communities, to communities both on- and offline, to influencers and policymakers. This does not necessarily equate to *actual* impact: For example, a policymaker may repeat an IO's message in a debate but have it rejected. Likewise, any actual impact may not necessarily be the result the IO wanted to achieve, given that the IO's intentions are unlikely to be known. For the purposes of this paper, the importance of the breakout scale is that it provides a way to approximate an operation's potential impact in close to real time. It can also be used to compare different operations according to a standard scale, which is based on observed incidents, rather than aggregated data flows or estimates of reach, impressions or viewing figures.

THE BREAKOUT SCALE: SIX CATEGORIES OF OPERATION

Using the concept of breakout moments as a guide, researchers can divide IOs into six categories on an ascending scale. Each category represents the influence operation at a specific moment in time, so operations can both rise up the scale and fall back down it. For example, the IRA of September 2014 (multiple platforms, little organic breakout) was very different from the IRA of late 2016 (multiple platforms and breakouts manifesting in policy impact), and this in turn was different from the IRA of November 2018 (multiple platforms and direct outreach, no breakout) or October 2019 (single platform, limited breakout).

The scale is actor-agnostic and can be used to compare influence operations, conspiracy theories, and a wide range of online activities. These can include deliberate disinformation efforts, the broader spread of misinformation tropes (such as coronavirus-related false information), and even the potential impact of official government communications compared with covert campaigns run by the same governments. It is especially intended for operational researchers, including at the social media platforms, as a rapid tool for reducing widely differing influence operations to a common scale, and thus enabling a prioritization of resources, and a greater degree of coordination, in the response.



Category One: one platform, no breakout

Category One operations exist on a single platform, and their messaging does not spread beyond the community at the insertion point.

The content may spread *within* that community, but it fails to reach new audiences. As such, it may reinforce that community's existing beliefs, but it has little opportunity to convert users in other communities, or to spread more broadly.

Politically themed clickbait and spam often falls into this category.[37] A Polish operation on Twitter in 2017, heavily astroturfed (meaning the actual creators were masked and the operation was made to look as if it had grassroots origins), that accused Polish political protesters of astroturfing, was a Category One. It generated 15,000 tweets in a few minutes, failed to catch on, and dropped back to zero within a couple of hours.[38] On Facebook, Iran's early attempts to interfere in the U.S. Republican Party's 2012 primaries and the 2014 independence referendum in Scotland were Category One efforts. They stayed on one platform and struggled to generate engagement even in the communities they were apparently targeting (supporters of Ron Paul or of Scottish independence).[39]

Category Two: one platform, breakout; many platforms, no breakout

Category Two operations either spread beyond the insertion point but stay on one platform, or feature insertion points on multiple platforms, but do not spread beyond them.

An example of the former is the IRA's IO iteration of 2019, dubbed "IRACopyPasta," which was almost exclusively active on Instagram.[40] This targeted multiple communities, including Black Lives Matter, Blue Lives Matter, LGBT, and Confederate groups, and some of its posts did achieve engagement from users who were not affiliated with the operation, indicating at least a degree of breakout on Instagram. However, this part of the operation remained on Instagram alone: Searches across other platforms did not reveal equivalent assets or repetitions of the same posts that could be reliably traced back to this source.

An example of the latter is the pro-Chinese government spam network "Spamouflage Dragon."[41] This posted political content mixed with spam (which we assume to have been a camouflage measure, hence the name) across YouTube, Facebook, and Twitter. It operated a substantial number of assets — in the hundreds across all three platforms — but all of the reactions to its posts came from other members of the same network. To date, no evidence has surfaced to suggest that substantial numbers of genuine users reacted to its posts. As such, the Spamouflage Dragon network existed on multiple platforms, but failed to break out of its insertion point on any of them.

Category Three: multiple platforms, multiple breakouts

Category Three influence operations feature insertion points and breakout moments on multiple platforms, but do not spread onto mainstream media.

More than most, Category Three is a transient category, in that influence operations seldom finish their lives as Category Threes: They tend to either remain stuck in the lower categories or accelerate onwards into Category Four. This is because stories that are substantial enough to break out of their insertion points and spread organically on multiple platforms are likely to draw the attention of tech and social media journalists, and thus to land in the traditional media as well.

The most notorious recent examples of Category Three efforts are conspiracy theories such as "Pizzagate" and "QAnon" before they were picked up and reported by the mainstream media. These efforts — which were deliberately deceptive works of fiction, but which appeared domestic in origin — started on fringe forums, notably 4chan, and picked up genuinely conspiracy-minded adherents on a range of social media platforms before they reached mainstream attention.

While usually ephemeral, Category Three is important in the overall scale, because it is the last category before an operation's story breaks out of the online medium entirely and makes it into traditional media. If researchers find an operation that they classify as Category Three, a timely exposure or other response will be crucial before the operation can break new ground.

Category Four: cross-medium breakout

Category Four operations manage to break out of the social media sphere entirely and are reported by the mainstream media, either as embedded posts or as reports.

The Iranian operation "Endless Mayfly" was an example of an IO that attained Category Four status: It created a fake website to run a false story about Qatar's preparation for the 2020 World Cup, which was briefly reported by Reuters.[42]

As noted above, the IRA achieved Category Four status on numerous occasions when tweets by its accounts were embedded into mainstream reporting. For example, in January 2017, the Los Angeles Times embedded tweets from two different IRA Twitter accounts into its reporting on the reaction to Starbucks's decision to offer jobs to refugees. With the embedded tweets, the IRA effectively supplanted the voice of the American alt-right in this article.[43] IRA persona Jenna Abrams was the focus of a brief article on "her" tweet about Kim Kardashian.[44] IRA persona @SouthLoneStar gained notoriety in the United Kingdom with an anti-Muslim tweet immediately after the London Bridge terrorist attack of March 2017, which was featured in coverage by some of the country's biggest tabloids.[45] Much of the coverage was hostile to the troll account, but the coverage itself exposed the account to a broad new audience.

Cross-medium breakout can work in both directions. In April 2017, the Russian state TV channel Zvezda ran the false claim that a Russian aircraft had disabled a U.S. Aegis cruiser by jamming it.[46] The false claim was based on an online article by a pro-Kremlin writer (breakout from electronic to traditional media). The Zvezda piece was then picked up by a range of blogs that spread it through conspiracy-minded communities on social media (breakout from traditional media in Russian to social media in English).

Category Five: celebrity amplification

Beyond mainstream media reporting, IOs reach Category Five status if celebrities amplify their messages — especially if they explicitly endorse them. This gives the information operators a powerful external validation, effectively attaching the celebrity's seal of approval and personal credibility to the operation's message.

President Donald Trump has been among those to give influence operations the boost to Category Five status. For example, in a campaign speech on September 28, 2016, he claimed that Google was "suppressing the bad news about Hillary Clinton."[47] The underlying theory had been debunked as early as June 2016,[48] but Kremlin outlet Sputnik ran a lengthy article on it on September 12,[49] and this

version of the story was amplified by pro-Trump outlets including Breitbart, which was likely Trump's source.[50] The overall story was not a Russian creation, but Sputnik played the pivotal role in amplifying it until it reached conservative American circles.

The term "celebrity" here is broadly defined: Politicians are not the only high-impact amplifiers who can boost influence operations and false claims. On one occasion, musician Roger Waters (formerly of Pink Floyd) falsely accused the White Helmets rescue group of being "fake," a claim spread by the Kremlin and the Assad regime.[51] On another, actor Woody Harrelson shared on Instagram the false claim that the 5G mobile phone network "may be exacerbating" the spread of COVID-19.[52]

Category Six: policy response; call for violence

An IO reaches Category Six if it triggers a policy response or some other form of concrete action, or if it includes a call for violence.[53] This is a (thankfully) rare category. Most Category Six influence operations are associated with hack-and-leak operations which use genuine documents to achieve their aim; they can also be associated with conspiracy theories or other operations that incite people to violence.

For example, the Russian hacking of the Democratic National Committee in 2016 was a Category Six operation because the subsequent leaks led to the resignation of several senior DNC staff, including chairwoman Debbie Wasserman Schultz.[54] The IRA's operation reached Category Six in May 2016 when it organized two conflicting demonstrations in Houston, Texas, and advised demonstrators to bring their arms. As noted above, there are legitimate reasons to question how effective the operation was in hindsight, but if operational researchers had identified it before the event, the danger of having two groups of armed and conflicting Americans facing off would have made it a matter of urgency.

Conspiracy theories can also tip over into Category Six if they come with the credible risk of violence. The "Pizzagate" theory that led an armed American to "self-investigate" a pizzeria in Washington, D.C.,[55] and the various anti-5G theories that led British arsonists to attack mobile-phone towers, fall into this category.[56] This is important to bear in mind, because such theories can seem so obviously false that it is easy to dismiss them out of hand. They can nevertheless cause real-world harm if they reach a susceptible audience.

IMPLICATIONS FOR RESPONSES

The primary purpose of this model is to allow operational researchers to situate the IO they are studying in the broader ecosystem of influence attempts, and to gauge its approximate impact relative to other known operations.

This should both allow responders to categorize and prioritize their responses to the IO and enable reporting of the operations on the basis of measurable and replicable evidence, thus reducing the danger of either panic or complacency.

The model also underscores the important role that influencers — both on social media and in society at large — play in providing IO with their breakout moments. This applies to journalists, who risk bringing an IO to new audiences if they fall for its messages; it also applies to politicians. Not least, it applies to celebrities and other public figures — all those who have a substantial audience, especially one that is not generally politically engaged, since these are the communities which are least likely to come across IOs via other routes.

Multiple IOs have shown how operators micro-target potential amplifiers via email, direct message, @-mentions, and other forms of direct outreach. Journalists are a favorite target, but the trade leaks, for example, also approached politicians and political activists directly.

It is influencers such as these who have the greatest potential to move IO out of the lower, barely noticed categories, into positions of prominence with substantial new audiences — and experience shows that the information operators know this. Politicians, journalists, and influencers should all beware direct outreach, and verify startling claims before they repeat them.

CONCLUSION

Not all influence operations are created equal. Many never spread beyond the platforms where they are planted, but a rare few can break out entirely and change the course of a political debate, at least temporarily. For operational researchers, the challenge is to determine which is which. The Breakout Scale is designed to allow such researchers to make and communicate that determination efficiently, based on criteria that are measurable, replicable and transparent.

But like influence operations, information consumers are created unequal: Mainstream journalists, politicians, and celebrities occupy a privileged and vulnerable position, because they have audiences far beyond the scope of the average citizen. Such influencers can make the difference between a weaponized leak or false story staying in the shadows and reaching a nationwide audience. The Breakout Scale is a reminder for influencers that they themselves can easily become the carrier, or the target, of influence operations. Their power to reach many people carries the responsibility to use that power with care.

REFERENCES

1 For the purposes of this paper, "disinformation" is defined as the deliberate spreading of false information, while "influence operations" are defined as efforts to influence public or political debate and decision-making processes that rely in part or in whole on covert activity.

2 "The conduct of information operations," (Washington, DC: Headquarters, U.S. Department of the Army, October 2018), https://fas.org/irp/doddir/army/atp3-13-1.pdf.

3 United States of America v. Internet Research Agency LLC et. al., Indictment 1:18-cr-00032-DLF, February 16, 2018, https://www.justice.gov/file/1035477/download.

4 Craig Malisow, "Hate Group Planning Islamic Library Protest Totally Doesn't Think They're a Hate Group," Houston Press, May 11, 2016, https://www.houstonpress.com/news/hate-group-planning-islamic-library-protest-totally-doesnt-think-theyre-a-hate-group-8393815.

5 Mike Glenn, "Dozens turn out to support Houston Muslims," *Houston Chronicle*, May 21, 2016, https://www.chron.com/news/houston-texas/houston/article/Dozens-turnout-to-support-Houston-Muslims-7926843.php#item-85307-tbla-10.

6 Ibid.

7 Erinn McQuagge, Rafael Linera, and Gregory Seese, "Effects-Based Psychological Operations Measures of Effectiveness: Measuring Change and Impact," (Arlington, VA: United States Department of Defense, March 2018), https://www.researchgate.net/publication/326957665_Effects-Based_Psychological_Operations_Measures_of_Effectiveness_Measuring_Change_and_Impact.

8 Andrew Jones, Jeremy Ellman, and Nanlin Jin, "An Application of Sentiment Analysis Techniques to Determine Public Opinion in Social Media," (Newcastle: Northumbria University, November 8, 2019), http://nrl.northumbria.ac.uk/41401/.

9 Sidney B Omongi Ochieng and Nanjira Sambuli, "Limitations of Sentiment Analysis on Facebook Data," *International Journal of Social Sciences and Information Technology* 2, no. 4 (June 15, 2016): 425-433, https://www.researchgate.net/publication/304024274_LIMITATIONS_OF_SENTIMENT_ANALYSIS_ON_FACEBOOK_DATA.

10 United States of America v. Internet Research Agency LLC et. al.

11 "Information Operations Directed At Hong Kong," Twitter Safety, August 19, 2019, https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html.

12 Kate Starbird, Ahmer Arif, and Tom Wilson, "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations," (Seattle: University of Washington, 2019), https://faculty.washington.edu/kstarbi/StarbirdArifWilson_DisinformationasCollaborativeWork-CameraReady-Preprint.pdf.

13 "Suspected Iranian Influence Operation Leverages Network of Inauthentic News Sites & Social Media Targeting Audiences in U.S., UK, Latin America, Middle East," FireEye Intelligence, August 21, 2018, https://www.fireeye.com/blog/threat-research/2018/08/suspected-iranian-influence-operation.html.

14 Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior in Thailand, Russia, Ukraine and Honduras," Facebook, July 25, 2019, https://about.fb.com/news/2019/07/removing-cib-thailand-russia-ukraine-honduras/.

15 Julian Borger, "US cuts funds for 'anti-propaganda' Iran group that trolled activists," *The Guardian*, May 31, 2019, https://www.theguardian.com/us-news/2019/may/31/us-cuts-funds-for-anti-propaganda-group-that-trolled-activists.

16 Nathaniel Gleicher, "Removing Coordinated Inauthentic Behavior in UAE, Egypt and Saudi Arabia," Facebook, August 1, 2019, https://about.fb.com/news/2019/08/cib-uae-egypt-saudi-arabia/.

17 Mike Isaac and Daisuke Wakabayashi, "Russian Influence Reached 126 Million Through Facebook Alone," *The New York Times*, October 30, 2017, https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html.

18 Twitter Safety, "Information Operations," Internet Research Agency archive of October 2018, https://transparency.twitter.com/en/reports/information-operations.html.

19 Kate Starbird, Ahmer Arif, and Tom Wilson, "Disinformation as Collaborative Work."

20 For example, one Iranian operation in early 2020 referred to the "U.S. navay" [sic] instead of the "U.S. Navy," allowing researchers to track each instance in which the original Iranian text was copied and pasted without editorial input.

21 Alicia Wanless, "We have a problem, but it isn't technology," Medium, May 29, 2019, https://medium.com/@lageneralista/we-have-a-problem-but-it-isnt-technology-c9163236767f.

22 Edmund L. Andrews, "How fake news spreads like a real virus," Stanford Engineering, October 9, 2019, https://engineering.stanford.edu/magazine/article/how-fake-news-spreads-real-virus.

23 Axelle Devaux et al, " Study on Media Literacy and Online Empowerment Issues Raised by Algorithm-Driven Media Services," Publications Office of the European Union, October 30, 2019, https://www.rand.org/pubs/external_publications/EP67990.html.

24 The term "insertion point" is preferred here over the traditional concept of the "target audience" to avoid the implication that the researcher can reliably gauge the IO's intent.

25 Ben Nimmo, "Robot Wars: How Bots Joined Battle in the Gulf," *Journal of International Affairs*, Columbia University, September 2018, https://jia.sipa.columbia.edu/robot-wars-how-bots-joined-battle-gulf.

26 Philip N. Howard, Bharath Ganesh, Dimitra Liotsiou, John Kelly, and Camille François, "The IRA, Social Media and Political Polarization in the United States, 2012-2018," (Oxford: University of Oxford, December 2018), https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf.

27 Alice Revelli and Lee Foster, "'Distinguished Impersonator' Information Operation That Previously Impersonated U.S. Politicians and Journalists on Social Media Leverages Fabricated U.S. Liberal Personas to Promote Iranian Interests," FireEye, February 12, 2020, https://www.fireeye.com/blog/threat-research/2020/02/information-operations-fabricated-personas-to-promote-iranian-interests.html.

28 Ben Nimmo, Graham Brookie, Nika Aleksejeva, Lukas Andriukaitis, Luiza Bandeira, Donora Barojan, Eto Buziashvili, Andy Carvin, Kanishk Karan, Iain Robertson, Michael Sheldon, "Operation Secondary Infektion," (Washington, DC: Atlantic Council, June 22, 2019), https://www.atlanticcouncil.org/wp-content/uploads/2019/08/Operation-Secondary-Infektion_English.pdf.

29 Ben Nimmo, "UK Trade Leaks," (New York: Graphika, December 2, 2019), https://graphika.com/tradeleaks.

30 The tone of the breakout may be positive, negative, or neutral. Some negative comment (such as exposure and debunking) may blunt the spread, but this is not universal: even negative comment can be beneficial to the IO in question if it introduces it or spreads its reputation to a new audience. ISIS' reputation for online savviness largely grew from the number of anti-ISIS articles that talked about it. See Brendan I. Koerner, "Why ISIS is winning the social media war," WIRED, March 2016, https://www.wired.com/2016/03/isis-winning-social-media-war-heres-beat/.

31 "Traditional media," in this context, refers to those media for which online production is not their primary function, such as newspapers, magazines, TV, and radio stations. An IO is considered to be cross-medium if its content features in the traditional media's primary form of publication — for example, an IRA tweet being quoted in a newspaper article or TV broadcast, not merely featured in the publication's Twitter feed.

32 Josephine Lukito and Chris Wells, "Most major outlets have used Russian tweets as sources for partisan opinion: study," Columbia Journalism Review, March 8, 2018, https://www.cjr.org/analysis/tweets-russia-news.php.

33 The threat of violence is a complex topic which intersects with, but is not coterminous with, information operations. For the purposes of this paper, which is aimed at the operational research community, any operation that calls for violence or appears intended to trigger it should be considered an imminent threat until proven otherwise. Retrospective analyses can, if necessary, downgrade the threat based on the outcome.

34 United States of America v. Viktor Borisovich Netyshko et. al., Indictment 1:18-cr-00215-ABJ, July 13, 2018, https://www.justice.gov/file/1080281/download.

35 Ibid, paragraph 47a.

36 Jonathan Martin and Alan Rappeport, "Debbie Wasserman Schultz to Resign D.N.C. Post," *The New York Times*, July 24, 2016, https://www.nytimes.com/2016/07/25/us/politics/debbie-wasserman-schultz-dnc-wikileaks-emails.html.

37 See for example Hannah Kozlowska, "A Ukrainian dad ran a giant, pro-Trump Facebook campaign with his 13-year-old kid," Quartz, September 25, 2019, https://qz.com/1716074/how-a-ukrainian-dad-built-a-huge-pro-trump-misinformation-machine/. The original reporting on this clickbait network claimed that it was "massive" and had "extraordinary reach," based on aggregated engagement numbers, but the network only existed on Facebook and its engagement appeared to be limited to one community, albeit a large one. This underlines the importance of measured and evidence-based reporting to avoid spreading further hysteria over the state of online operations.

38 Ben Nimmo, "Polish Astroturfers Attack… Astroturfing," DFRLab, July 23, 2017, https://medium.com/dfrlab/polish-astroturfers-attack-astroturfing-743cf602200.

39 Ben Nimmo, C. Shawn Eib, Léa Ronzaud, Rodrigo Ferreira, Thomas Lederer, and Melanie Smith, "Iran's Broadcaster: Inauthentic Behavior," (New York: Graphika, May 5, 2020), https://graphika.com/reports/irans-broadcaster-inauthentic-behavior/.

40 Camille François, Ben Nimmo, and C. Shawn Eib, "The IRA CopyPasta Campaign," (New York: Graphika, October 21, 2019), https://graphika.com/reports/copypasta/. Facebook reported one account on Facebook itself, but this does not appear to have been active.

41 Ben Nimmo, C. Shawn Eib, and L. Tamora, "Spamouflage Dragon," (New York: Graphika, September 25, 2019), https://graphika.com/reports/spamouflage/.

42 Gabrielle Lim et al, "Burned After Reading: Endless Mayfly's Ephemeral Disinformation Campaign," Citizen Lab, May 14, 2019, https://citizenlab.ca/2019/05/burned-after-reading-endless-mayflys-ephemeral-disinformation-campaign/.

43 Nina Agrawal, "Supporters and opponents of Trump's refugee ban take to social media to put pressure on companies," *LA Times*, January 30, 2017, https://www.latimes.com/nation/la-na-boycott-starbucks-twitter-20170130-story.html. The embeds were from IRA accounts @TEN_GOP and @Pamela_Moore13.

44 Desiree O., "This Tweeter's PERFECT Response to Kim K's Naked Selfie Will Crack You Up," Brit + Co, March 7, 2016, https://www.brit.co/tweeter-kim-kardashian-naked/.

45 Alex Hern, "How a Russian 'troll soldier' stirred anger after the Westminster attack," *The Guardian*, November 14, 2017, https://www.theguardian.com/uk-news/2017/nov/14/how-a-russian-troll-soldier-stirred-anger-after-the-westminster-attack.

46 Ben Nimmo, "Russia's Fake 'Electronic Bomb'," DFRLab, May 8, 2017, https://medium.com/dfrlab/russias-fake-electronic-bomb-4ce9dbbc57f8.

47 Nick Corasaniti, "Donald Trump Pushes Debunked Theory That Google Suppressed Rival's Bad News," *The New York Times*, September 28, 2016, https://www.nytimes.com/2016/09/29/us/politics/google-trump-clinton.html.

48 Dan Evon, "Does This Video Document Google Manipulating Searches for Hillary Clinton?," Snopes, June 10, 2016, https://www.snopes.com/fact-check/google-manipulate-hillary-clinton/.

49 "SPUTNIK EXCLUSIVE: Research Proves Google Manipulates Millions To Favor Clinton," Sputnik, September 12, 2016, https://sputniknews.com/us/201609121045214398-google-clinton-manipulation-election/.

50 Jack Hadfield, "Report: Google search bias protecting Hillary Clinton confirmed in experiment," Breitbart, September 13, 2016, https://www.breitbart.com/tech/2016/09/13/hillary-google-bias-confirmed-experiment/. The Breitbart article named Sputnik as its source.

51 "Roger Waters claims Syria's White Helmets a 'fake organization'," *Times of Israel*, April 16, 2018, https://www.timesofisrael.com/roger-waters-claims-syrias-white-helmets-a-fake-organization/.

52 The post is preserved at https://archive.is/2XI3s.

53 For a discussion on dangerous speech and the criteria which can be used to identify it, see Susan Benesch's Dangerous Speech Project at https://dangerousspeech.org/faq/.

54 Jonathan Martin and Alan Rappeport, "Debbie Wasserman Schultz to Resign D.N.C. Post."

55 Camila Domonoske, "Man Fires Rifle Inside D.C. Pizzeria, Cites Fictitious Conspiracy Theories," NPR, December 5, 2016, https://www.npr.org/sections/thetwo-way/2016/12/05/504404675/man-fires-rifle-inside-d-c-pizzeria-cites-fictitious-conspiracy-theories.

56 "Conspiracy theorists burn 5G towers, incorrectly linking them to coronavirus," Associated Press, April 21, 2020, https://www.pennlive.com/coronavirus/2020/04/conspiracy-theorists-burn-5g-towers-incorrectly-linking-them-to-coronavirus.html.

ABOUT THE AUTHOR

Ben Nimmo is the Head of Investigations at Graphika, a New York-based social media analytics company. Since 2014, he has been studying and exposing influence operations across multiple platforms and geographical areas. He specializes in investigating cross-platform activity. He previously worked as a journalist and a press officer at NATO. He has bachelor's and master's degrees from Cambridge University, and speaks a number of languages, including French, Russian, and Latvian.

ACKNOWLEDGEMENTS

The author would like to thank all the investigators whose work contributed to the case studies on which this paper is based, especially the teams at Graphika and the DFRLab. Particular thanks are due to Alicia Wanless and Camille François for their expertise and thoughtful feedback on the early drafts.

Caroline Klaff, Sarah Reed, and Anna Newby edited this paper, and Rachel Slattery provided layout.