The Instigators and Targets of Organised Social Media Manipulation:

Global Index 2022

Hannah Bailey and Philip N. Howard

EXECUTIVE SUMMARY

Who are the actors responsible for manipulating information online? And who are these actors targeting with this manipulated content? To understand the sources and targets of online information operations, we analyze 89,104 suspended Twitter accounts and 65,659 Facebook and Instagram accounts taken down between 2017-2021. We find:

- Many of the countries that are responsible for instigating information operations on social media platforms are also recipients of these operations. Of the 25 countries most responsible for initiating information operations, 15 of these countries are also among the 25 countries that are most heavily targeted by information operations on Facebook and Instagram. These 15 countries are Brazil, Egypt, Georgia, Indonesia, Iran, Iraq, Mexico, Myanmar, Nicaragua, Pakistan, Russia, Sudan, Thailand, Ukraine, and the United States of America.
- China is particularly conspicuous for its role as the dominant instigator of information operations, and the lack of operations on Facebook and Instagram targeted toward its domestic population. This is likely because Facebook and Instagram are largely inaccessible to China's domestic online audiences.
- Information operations detected by Twitter predominantly target non-English language audiences. Only 6% of posts from suspended inauthentic accounts are in English. Turkish and Arabic speaking audiences are heavily targeted by information operations. Of the total 257 million inauthentic Tweets detected by Twitter, 27% are in Turkish and 30% are in Arabic. In fact, Arabic, Spanish and Turkish together account for 64% of all the content removed by Twitter for violating its guidelines.

1. INTRODUCTION

Just as social media has dramatically increased the widespread dispersion of information, it has also dramatically increased the ease with which this information may be manipulated. But who are the actors that lie at the heart of this manipulated information? Because information has become a critical resource of the 21st century, it is important to identify both the instigators and the targets of this information manipulation.

In this report, we focus on "information operations". We adopt the definition for these operations as given by Meta (Facebook and Instagram) that is, social media campaigns "organized [by] actors (governments or non-state actors) to distort domestic or foreign political sentiment, most frequently to achieve a strategic and/or geopolitical outcome".[1] The methods used by information operations can include spreading inaccurate or misleading information, or inauthentically amplifying particular social media posts.

Social media firms frequently take down—or suspend—accounts as a means to disrupt information operation networks. Both Twitter and Meta regularly suspend networks of accounts on their respective platforms for conducting information operations. Since 2017, Meta has publicly released data from these suspended networks. [2] Twitter followed suit in 2018. [3] There are, however, differences in practices. Twitter releases data which include the country from which the accounts operated, as well as account level data, such as the messages posted by these accounts and the languages used in messages. In contrast, Meta release data that include the country of origin for each account, the country or region targeted and the number of accounts in the information operations network. Thus, Meta data does not include account level data and Twitter data does not explicitly identify the targets of the information operation. Using data from both Meta and Twitter allows us to leverage the maximum amount of information on both the user accounts and their regional profiles.

Here, we analyze information operation takedown data from Twitter, Facebook and Instagram, over the period from 2017-2021. Notably, these include accounts existing prior to this period. Specifically, we analyze 89,104 suspended Twitter accounts that were taken down between 2018-2021. This includes over 257 million Tweets posted before suspension, with a total user engagement with these Tweets of over 754 million using 74 unique languages. We also analyze 65,659 Facebook and Instagram accounts taken down between 2017-2021. In total 68 countries were targeted by these Facebook and Instagram accounts, and 79 countries instigated these accounts.

With these data, we identify geographic regions that are disproportionately targeted by online information operations, as well as countries that are disproportionately responsible for these information operations.

This report builds on our previous "cyber troop" investigations. Since 2016 we have monitored the global rise in information operations, and the evolving techniques used by the actors conducting these campaigns. In recent reports, we found that more information operations are being outsourced to private firms, [4] and that authoritarian regimes are using information operations as a tool to suppress human rights. [5]

While other reports have examined individual information operation networks detected by Twitter and Meta, [6][7][8][9] we uniquely combine these data to provide a comprehensive overview of the instigators and targets of information operations on the largest global social media platforms.

To summarize in advance, our findings are as follows. First, many of the countries that are responsible for instigating information operations on social media platforms are also recipients of these operations. Of the 25 countries most responsible for initiating information operations, 15 of these countries are also among the 25 countries that are most heavily targeted by information operations on Facebook and Instagram. These 15 countries are Brazil, Egypt, Georgia, Indonesia, Iran, Iraq, Mexico, Myanmar, Nicaragua, Pakistan, Russia, Sudan, Thailand, Ukraine, and the United States of America.

Second, China is particularly conspicuous for its role as the dominant instigator of information operations, and lack operations on Facebook and Instagram targeted toward its domestic population. This is likely because Facebook and Instagram are largely inaccessible to China's domestic online audiences.

Third, we find that information operations detected by Twitter predominantly target non-English-language audiences. Only 6% of posts from suspended inauthentic accounts are in English. In contrast, Turkish and Arabic speaking audiences are heavily targeted. Of the total 257 million inauthentic Tweets detected by Twitter, 27% are in Turkish and 30% are in Arabic.

2. THE INSTIGATORS OF INFORMATION MANIPUATION ON FACEBOOK, INSTAGRAM AND TWITTER

3.3 The Countries Instigating Information Operations on Facebook, Instagram and Twitter

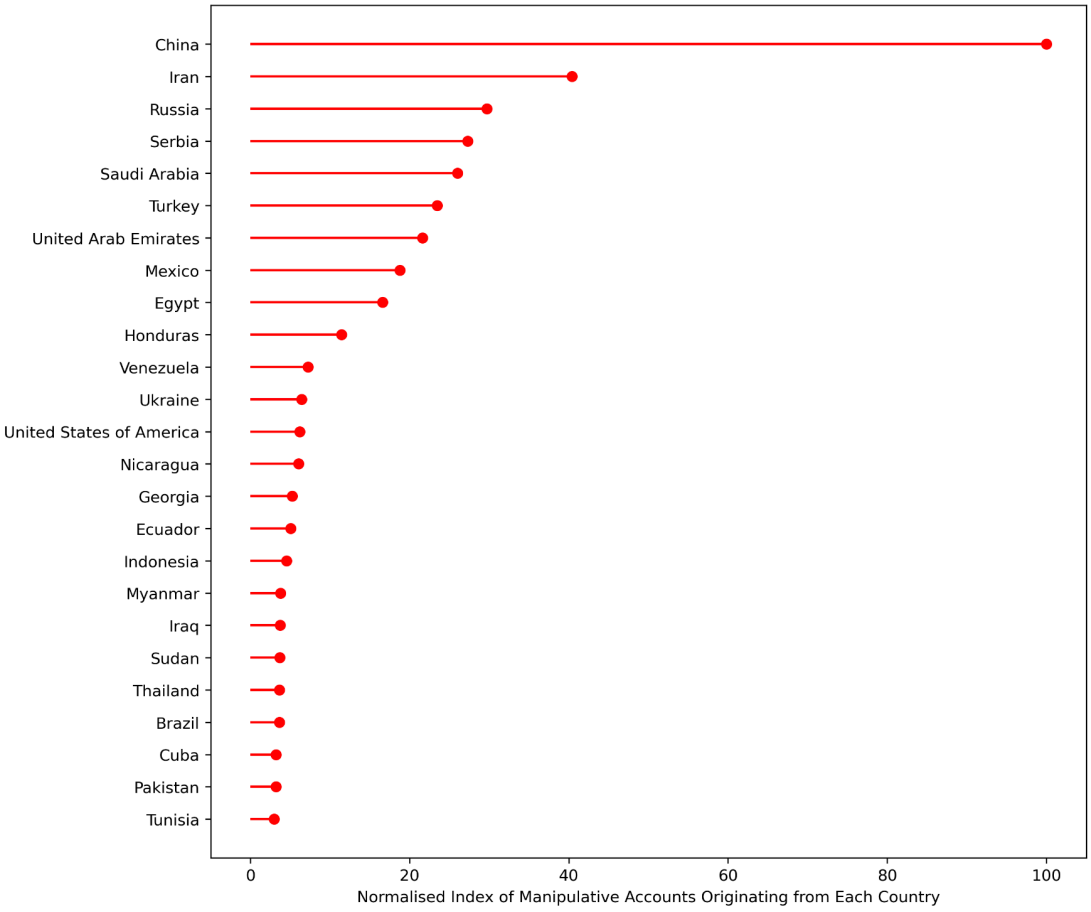
As a first cut into our data, we seek to distinguish information operations by the different countries. Specifically, we focus broadly on two types of actors: (1) countries that instigate information manipulation; and (2) countries that are targeted with information manipulation.

To assess the first category of countries we use both Meta (Facebook/Instagram) and Twitter data. The total number of accounts across the set of countries ranges from 0 to 31,347. To better represent the distribution across countries, we rescale the data to range between 0-100, using a min-max normalization. The normalized data are illustrated in *Figure 1*.

The conspicuous dominance of China is clearly seen in this figure, with a normalized score of 100 and an N of 31,347. China accounts for more than twice the second ranked country, Iran. Broadly speaking, China might arguably consist as its own top category, with a second grouping in the mid-range. This mid-range group of countries consists of Iran, Russia, Serbia, Saudi Arabia, Turkey, United Arab Emirates, Mexico, Egypt and Honduras. Within this group, Iran ranks highest with a normalized score of 40.4 (N of 12,664) and Honduras is the lowest with a score of 11.4 (N of 3,580).

A third category of country within the top 25 ranges in normalized scores from about 7 to 3 (N from 2,270 to 933). The remaining countries in our dataset fall below an N of 933.

Figure 1: Top 25 Countries Instigating Information Manipulation on Facebook, Instagram and Twitter



Source: Account takedown data from Meta and Twitter.

Note: Country-level data are normalised using min-max normalisation between 0-100.

3.1 The Countries Targeted by Information Operations on Facebook and Instagram

Our second analysis is of those countries that were targeted by information operations. Importantly, this also includes countries in which information operations were also targeted at the domestic population. A second note of caution is that Twitter does not provide data on recipients of information operations, and so these data include only data from Facebook and Instagram.

In *Figure 2* we use the same min-max normalization process and again identify the top 25 countries that were targeted by information manipulation.

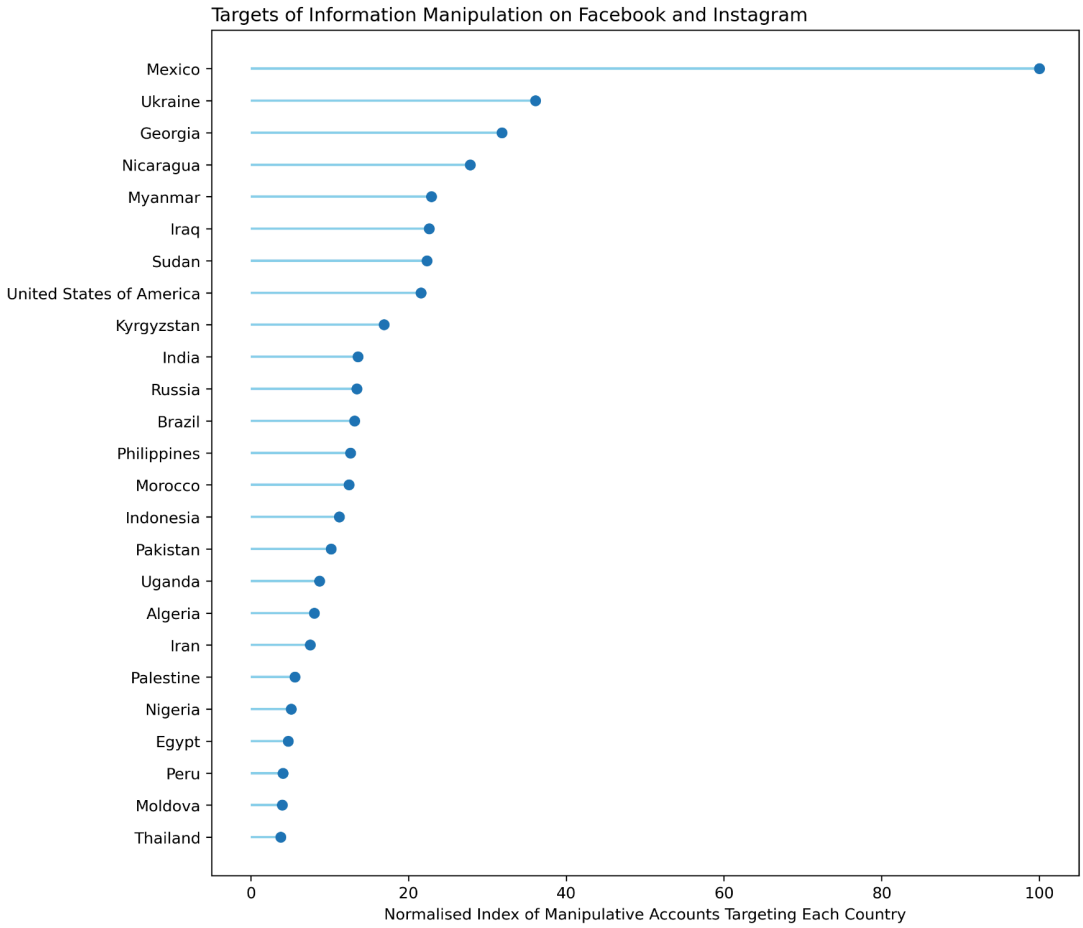
Similar to *Figure 1*, in *Figure 2* one country exhibits dominance, and in this case that is Mexico, normalized at a score at 100 with an N of 5,165. However, the number of accounts targeting Mexico is much smaller than the number of accounts from China that were instigating information operations.

Again, we group the countries into three categories with the dominant country in this case, Mexico comprising the top category. A mid-range group of countries consists of Ukraine, Georgia, Nicaragua, Myanmar, Iraq, Sudan, and the United States of America. Within this group, Ukraine ranks highest with a normalized score of 36 (N of 1,863). And the United States of America is the lowest with a score of 21.5 (N of 1,113).

For *Figure 2,* a third category of country within the top 25 ranges in normalized scores from about 17 to 4 (N from 871 to 196). The remaining countries in our dataset fall below an N of 196.

Intriguingly, *Figure 2* shows that many of the countries that instigate information operations (*Figure 1*) are also recipients of these operations. Of the top 25 countries which initiate information operations (*Figure 1*), 15 also appear in the top 25 targeted countries (*Figure 2*). These 15 countries are Brazil, Egypt, Georgia, Indonesia, Iran, Iraq, Mexico, Myanmar, Nicaragua, Pakistan, Russia, Sudan, Thailand, Ukraine, and the United States of America. Notably, while China appeared as the dominant instigator of information operations, it does not appear to be heavily targeted. This is likely because domestic audiences in China do not have access to Facebook and Instagram, so any information operations targeted toward this audience would not appear on these platforms.

Figure 2: Top 25 Targets of Information Manipulation on Facebook and Instagram



Source: Account takedown data from Meta.

Note: Country-level data are normalised using min-max normalisation between 0-100.

3. LANGUAGES TARGETED BY INFORMATION OPERATIONS

Information operations invariably target a particular geographic region or online community by posting messages in the language used by that community. By examining the language used in an information operation, we can therefore infer the region or online community that was targeted by that operation.

Figure 3 shows the five most prevalent languages used by information operations on Twitter. This figure plots the number of Tweets posted in each language by accounts that were later suspended by Twitter for conducting information operations. Notably, the representation of these languages is contingent upon Twitter identifying and suspending accounts in these languages. That is, there is no assurance by Twitter that all languages are equally scrutinized, and thereby subjected to account suspension.

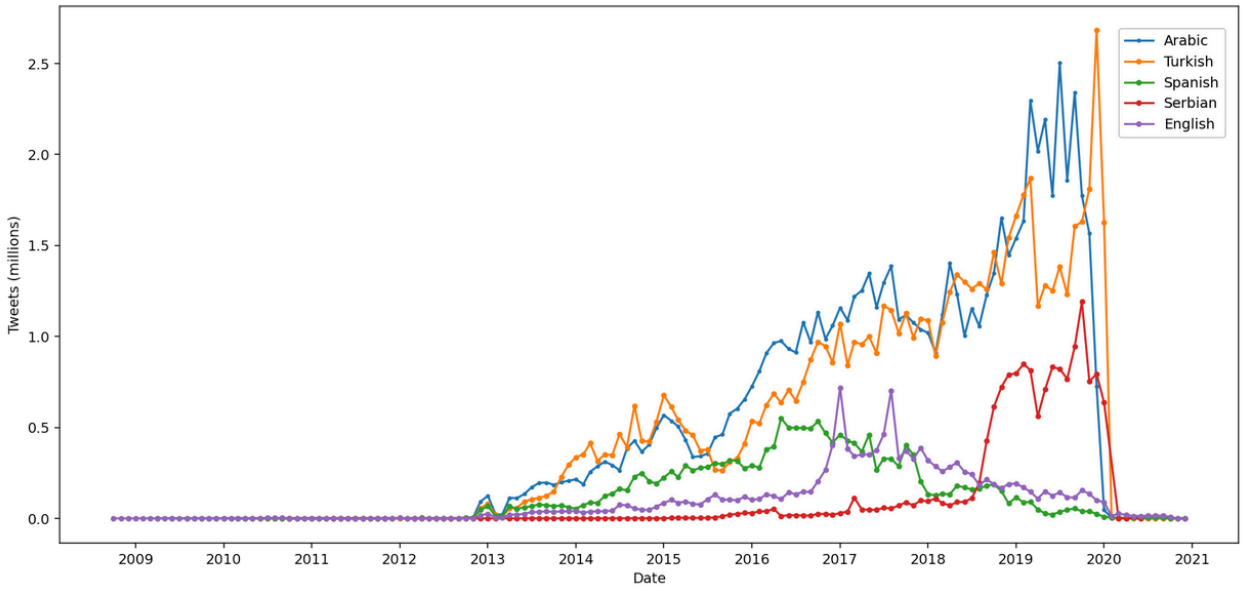
Bearing in mind the above caveat, *Figure 3* illustrates that Arabic is the language used most frequently by information operations on Twitter, with a total of 74 million posts over a fourteen year period. Turkish is the second most used language, with 68.5 million Tweets, Spanish is third at 17.8 million and Serbian is fourth with 15 million. Although it is the world's most spoken language, for the data we analyze, only 14 million Tweets by information operations were in English.

Tweets in Turkish and Arabic show a steady increase between 2013 and early 2020. For both languages, the number of Tweets by information operations reached over 2.5 million per month between 2019 and early 2020. Notably, the number of Tweets from all languages falls dramatically in 2020 for the simple reason that Twitter had suspended these accounts. Hence, *Figure 3* plots the activity of accounts that were taken down at some point between 2018 and 2021. We thus observe the impact of these takedowns most clearly in early 2020, where 52,660 accounts were suspended between March and June.

Tweets in Serbian show a sharp increase in activity between late 2018 and early 2020. This coincides with the 2018-2020 protests in Serbia against the ruling Serbian Progressive Party. [10] In its takedown announcement, Twitter noted that the inauthentic operation had been spreading pro-government propaganda, as well as attacking the protestors and political opponents. [11]

The broad takeaway from this is that, perhaps counter intuitively, English is not the language in which the vast majority of Twitter-detected information operation content is published. Instead, communities that use languages like Arabic, Turkish, Spanish and Serbian are the primary recipients of information operations. In total, only 6% of all information operation Tweets from suspended accounts are in English. In contrast, 27% such Tweets are in Turkish and 30% in Arabic. It is important to note, however, that 14% of Tweets in this dataset are also labelled 'Undetermined Language' by Twitter. Once again, we also recognize that these conclusions are contingent upon the data available from Twitter, which may of course be subject to the observational bias . That is, Twitter may happen to find information operations in languages that it happens to be actively monitoring.

Figure 3: Information Operations Tweets in the Five Most Prevalent Languages



Source: Authors' calculations based on Twitter's information operation take-down reports between 2018 – 2021.

Notes: This figure plots the number of Tweets by information operations in each language per month. A table with the total number of inauthentic tweets for the ten most prevalent languages can be found in Appendix A.1.

4. CONCLUSION

The analysis of countries in *Figures 1* and *2* reveals one important finding: many of the same countries that instigate information operations are also recipients of these operations. For instance, 15 of the top 25 countries which initiate information operations (*Figure 1*) also appear in the top 25 targeted countries (*Figure 2*). These 15 countries are Brazil, Egypt, Georgia, Indonesia, Iran, Iraq, Mexico, Myanmar, Nicaragua, Pakistan, Russia, Sudan, Thailand, Ukraine, and the United States of America. The one conspicuous country which appears only in *Figure 1,* and as the dominant instigator of information operations, is China.

From *Figure 3* we find that information operations detected by Twitter predominantly target non-English-language audiences, in particular Arabic and Turkish speaking communities. Only 6% of posts from suspended inauthentic accounts are in English. In contrast, 27% of inauthentic Tweets detected by Twitter are in Turkish and 30% in Arabic.

This report offers the foundation for further research. In particular, our findings raise questions which extend beyond the remit of this short report. For instance, and bearing in mind the limitations of Twitter's language-based data collection methods, why do we find that non-English languages like Arabic and Turkish are the most prominent languages for information operations Tweets?

And, finally, why do countries which instigate information operations target their own domestic populations—and do the reasons for this differ from country to country? While we do not conjecture on the answers to these questions, our study provides a launching pad for important research questions.

REFERENCES

[1] J. Weedon, W. Nuland and A. Stamos, " Information Operations and Facebook," Apr. 2017. [Online]. Available: https://i2.res.24o.it/pdf2010/Editrice/ILSOLE24ORE/ILSOLE24ORE/Online/_Oggetti_Embedded/Documenti/2017/04/28/facebook-and-information-operations-v1.pdf

[2] Meta, "Meta | Coordinated Inauthentic Behaviour," 2022. https://about.fb.com/news/tag/coordinated-inauthentic-behavior/

[3] Twitter, " Twitter | Information Operations," 2022. https://transparency.twitter.com/en/reports/information-operations.html

[4] S. Bradshaw, H. Bailey and P. N. Howard, " Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation," Oxford, UK: Programme on Democracy and Technology, Oxford University, 2021. [Online]. Available: https://demtech.oii.ox.ac.uk/research/posts/industrialized-disinformation/

[5] S. Bradshaw and P. N. Howard, " The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation," Oxford, UK: Programme on Democracy and Technology, Oxford University, 2019. [Online]. Available: https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1209&context=scholcom

[6] S. Grossman, H. Khadija, R. DiResta, T. Kheradpir, and C. Miller, " Blame it on Iran, Qatar, and Turkey: An analysis of a Twitter and Facebook operation linked to Egypt, the UAE, and Saudi Arabia," Stanford Internet Observatory, Stanford University, 2020. [Online]. Available: https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/20200402_blame_it_on_iran_qatar_and_turkey_v2_0.pdf

[7] J. Uyheng, et al., "Interoperable pipelines for social cyber-security: assessing Twitter information operations during NATO Trident Juncture 2018," *Computational and Mathematical Organization Theory,* vol. 26, no. 4, pp. 465-483, 2020.

[8] X. Guo, and S. Vosoughi, "A Large-Scale Longitudinal Multimodal Dataset of State-Backed Information Operations on Twitter," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022.

[9] M. Schliebs, H. Bailey, J. Bright, & P. N. Howard, "China's public diplomacy operations: understanding engagement and inauthentic amplifications of PRC diplomats on Facebook and Twitter," Oxford, UK: Programme on Democracy and Technology, Oxford University, 2021. [Online]. Available: https://demtech.oii.ox.ac.uk/china-public-diplomacy-report

[10] BBC, " Serbia protests: Thousands march against President Vucic," *BBC*, Jan. 05 2019. [Online]. Available: https://www.bbc.co.uk/news/world-europe-46772500.

[11] D. Bush, " “Fighting Like a Lion for Serbia”: An Analysis of Government-Linked Influence Operations in Serbia," Stanford Internet Observatory, Stanford University, 2020. [Online]. Available: https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/serbia_march_twitter.pdf

ABOUT THE PROJECT

The Programme on Technology and Democracy investigates the use of algorithms, automation, and computational propaganda in public life. This programme of activity is backed by a team of social and information scientists eager to protect democracy and put social data science to work for civic engagement. We are conducting international fieldwork with the political consultants and computer experts who are commissioned to activate or catch information operations. We are building original databases of incidents and accounts involved in such activities, and we use our knowledge to make better tools for detecting and ending interference with democracy. We engage in "real-time" social and information science, actively disseminating our findings to journalists, industry, and foreign policy experts. Our network of experts helps civil society, industry, government, and other independent researchers develop a better understanding of the role of technology in public life.

AUTHOR BIOGRAPHIES

Hannah Bailey is a Researcher at the Oxford Internet Institute's Programme on Democracy and Technology, with a focus on social data science. Her research examines the PRC's use of state-sponsored digital disinformation. In particular, she focusses on the effect of the PRC's digital disinformation campaigns on international audiences by assessing how they interact with this disinformation. She holds a BSc in Politics and Philosophy from the London School of Economics, as well as two MScs, in Contemporary Chinese Studies, and in the Social Science of the Internet, both from Oxford University. She has also studied Mandarin at Fudan University (Shanghai). She tweets from @Hannah_LSBailey.

Philip N. Howard is a professor and writer. He teaches at the University of Oxford and directs the Programme on Democracy and Technology. He writes about information politics and international affairs, and he is the author of ten books, including The Managed Citizen, Pax Technica, and Computational Propaganda. He has won multiple best book awards, and his research and commentary writing has been featured in the New York Times, Washington Post, and many international media outlets. Foreign Policy magazine named him a "Global Thinker" for 2018 and the National Democratic Institute awarded him their "Democracy Prize" for pioneering the social science of fake news. He has testified before the US Senate, UK House of Parliament, and European Commission on the causes and consequences of fake news and misinformation. His latest book is Lie Machines: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives. He tweets from @pnhoward.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Ford Foundation, Luminate, and Craig Newmark Philanthropies. Furthermore, some authors of this work were supported by the Economic and Social Research Council (UKRI Grant Number 2260175)

Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Oxford, the Oxford Internet Institute, or our funders.

APPENDICES

A.1 Supporting Figures

Figure 4: Engagement per Information Operation Tweet

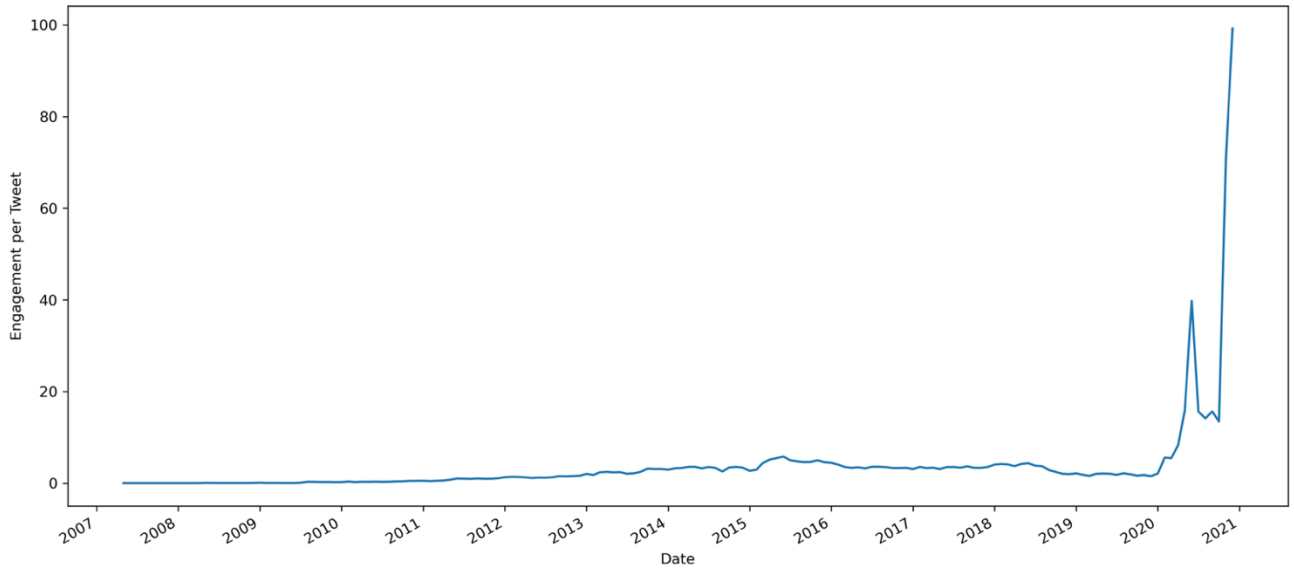
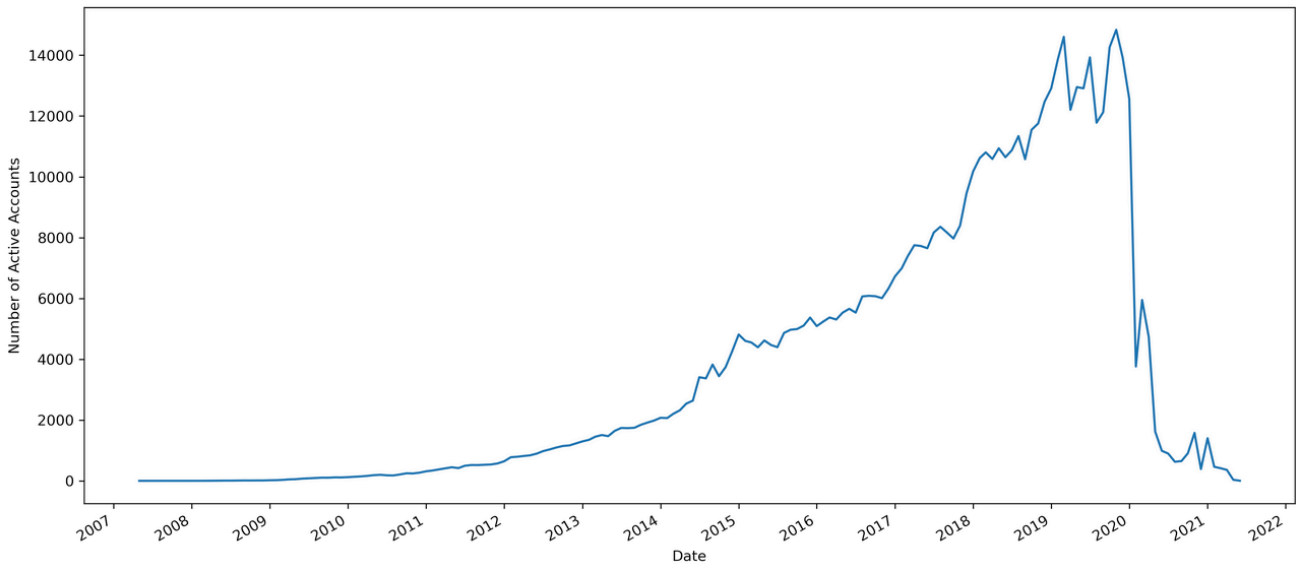


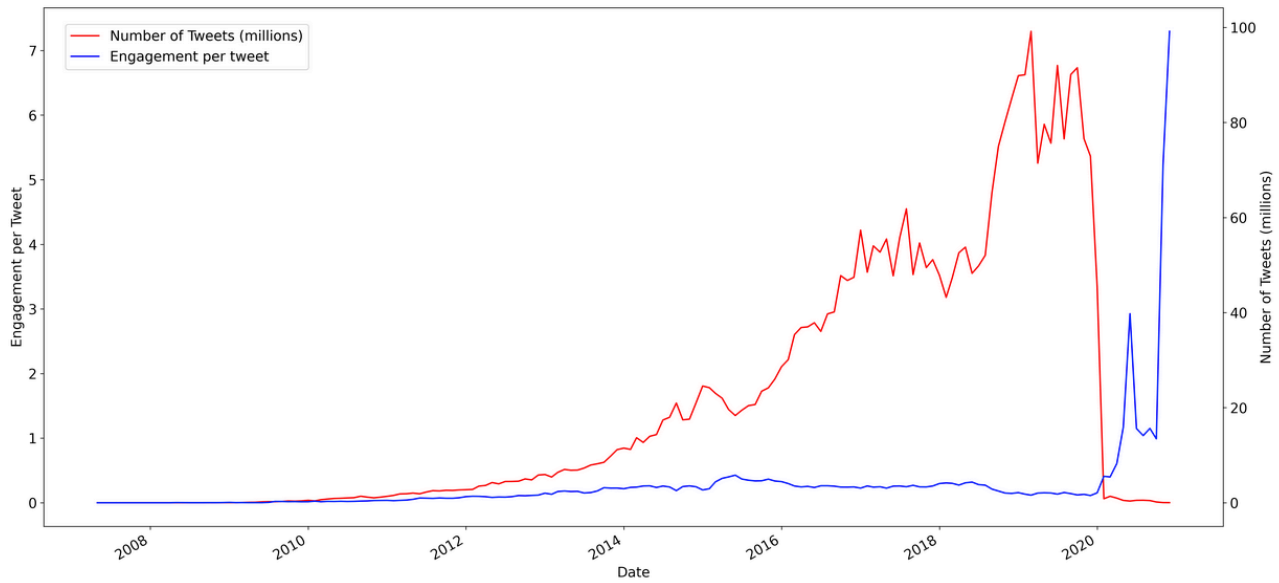
Figure 5: Twitter Takedown Activity, by Number of Active Accounts



Source: Twitter Information Operations Takedown Data, 2018 – 2022.

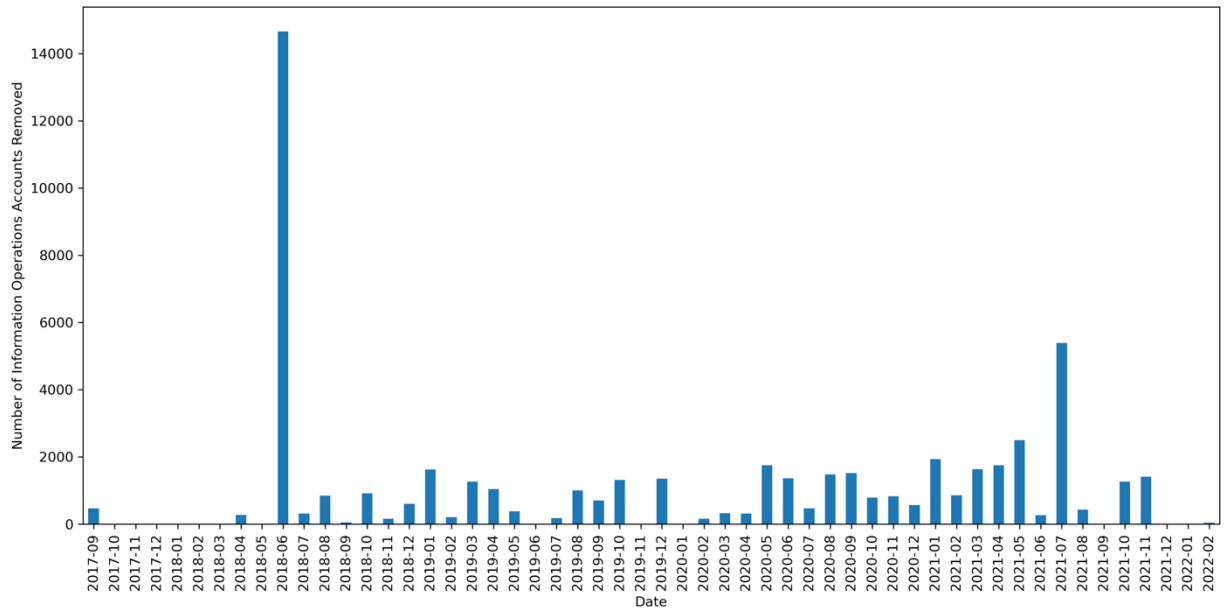
Notes: Number of active accounts are calculated per month. While Twitter only began taking down information operation accounts in 2018, these accounts were active before this time.

Figure 6: Aggregate Number of Information Operation Tweets, and the Average Engagement per Tweet, 2017-2021



Source: Twitter Information Operation take-down reports between 2018 – 2022.

Figure 7: Facebook Takedown Activity, by Number of Accounts



Source: Meta Information Operation take-down reports between 2017 – 2022.

Table 1: Total Number of Inauthentic Tweets per Language for the Ten Most Prevalent Languages, 2008 - 2021

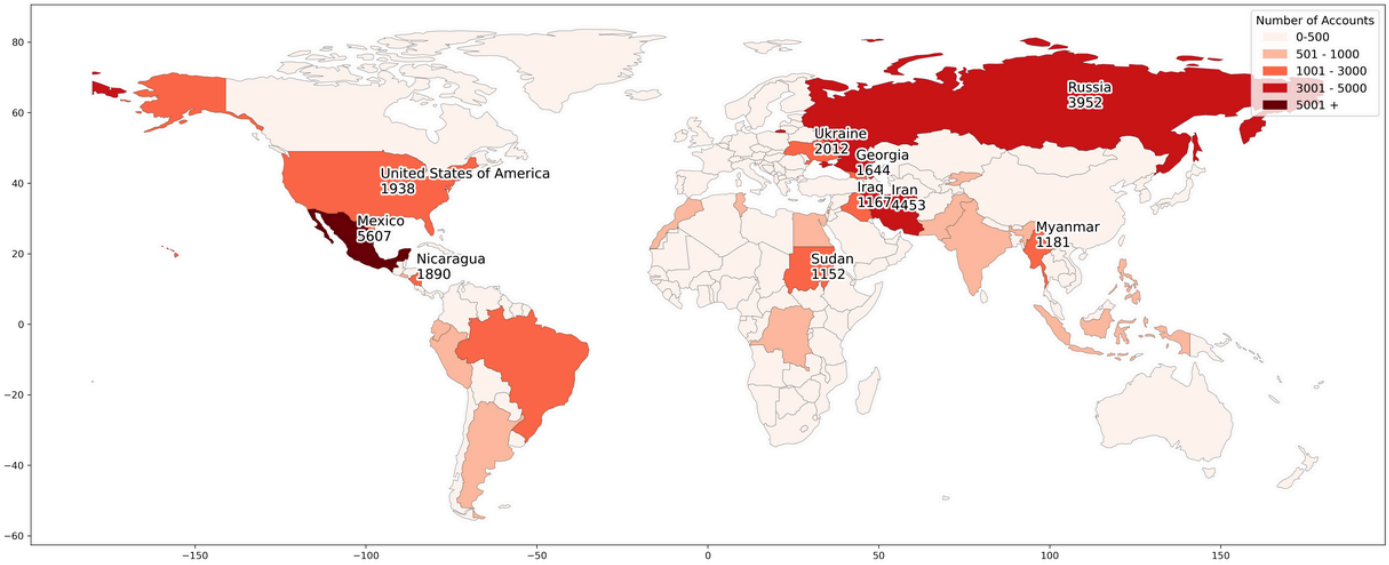
Language	Number of Tweets
Arabic	74,008,622
Turkish	68,534,113
Undetermined	35,752,207
Spanish	17,893,133
Serbian	15,042,635
English	14,025,526
Indonesian	7,399,012
Russian	5,555,734
Urdu	2,062,934
Persian	2,011,891

Source: Twitter Information Operation take-down reports between 2018 – 2022.

Notes: The "Undetermined" category is the Twitter category label given where the language used in the Tweet cannot be determined.

A.2 Supporting Figures

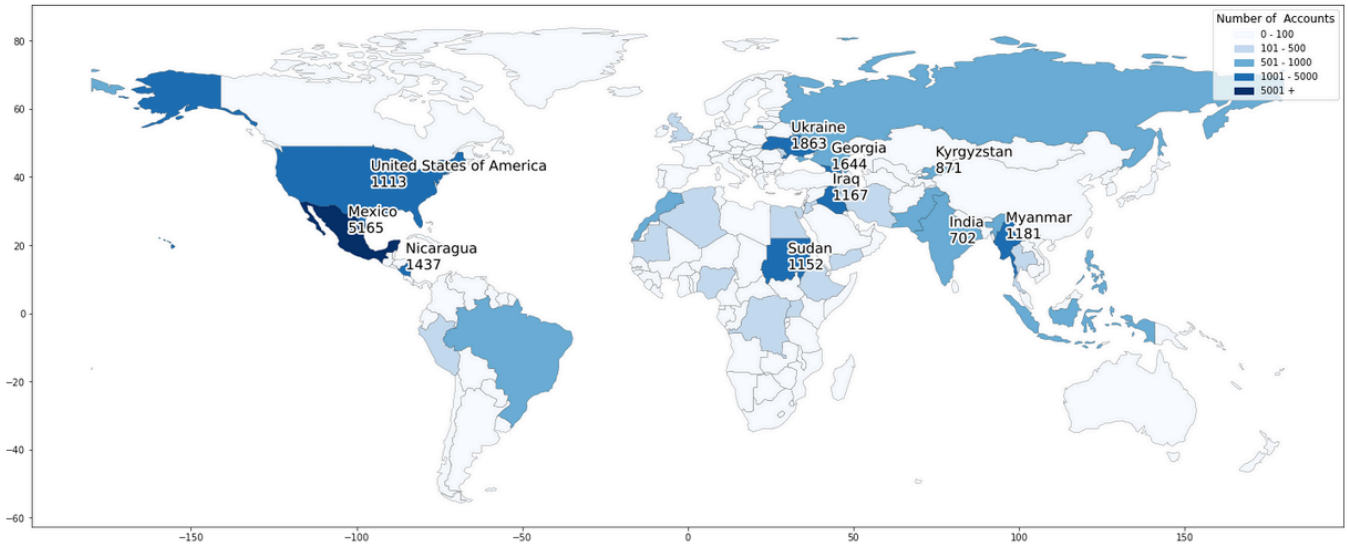
Figure 8: Heatmap of Countries Responsible for Instigating Information Operations on Facebook and Instagram, 2017-2022



Source: Meta Information Operation take-down reports between 2017 – 2022.

Notes: Graph displays the number of information operation accounts detected and taken down by Meta for each country between 2017 and 2022. Where Meta provides only a region responsible for the operation, these accounts are divided between the countries that comprise that region.

Figure 9: Heatmap of Information Operations on Facebook and Instagram Targeting Domestic Audiences, 2017-2022



Source: Meta Information Operation take-down reports between 2017 – 2022.

Notes: Graph displays the number of information operation accounts detected and taken down by Meta for each country between 2017 and 2022. Where Meta provides only a region or language targeted by an operation, these accounts are divided between the countries that comprise that region or for which the targeted language is the primary spoken language.