

# Characterization and Comparison of Russian and Chinese Disinformation Campaigns



David M. Beskow and Kathleen M. Carley

**Abstract** While substantial research has focused on social bot classification, less computational effort has focused on repeatable bot characterization. Binary classification into “bot” or “not bot” is just the first step in social cybersecurity workflows. Characterizing the malicious actors is the next step. To that end, this paper will characterize data associated with state sponsored manipulation by Russia and the People’s Republic of China. The data studied here was associated with information manipulation by state actors, the accounts were suspended by Twitter and subsequently all associated data was released to the public. Of the multiple data sets that Twitter released, we will focus on the data associated with the Russian Internet Research Agency and the People’s Republic of China. The goal of this paper is to compare and contrast these two important data sets while simultaneously developing repeatable workflows to characterize information operations for social cybersecurity.

**Keywords** Bot characterization · Social cybersecurity · Disinformation · Information operations · Strategic competition · Propaganada · Exploratory data analysis · Internet memes

## 1 Introduction

State and non-state actors leverage information operations to create strategic effects in an increasingly competitive world. While the art of influence and manipulation dates back to antiquity, technology today enable these influence operations at a scale and sophistication unmatched even a couple decades ago. Social media platforms have played a central role in the rise of technology enabled information warfare. As state and non-state actors increasingly leverage social media platforms as central to

---

D. M. Beskow (✉) · K. M. Carley  
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA  
e-mail: [dbeskow@andrew.cmu.edu](mailto:dbeskow@andrew.cmu.edu); [kathleen.carley@cs.cmu.edu](mailto:kathleen.carley@cs.cmu.edu)

**Table 1** List of Datasets that Twitter has released in association of state sponsored information manipulation

Year-Month	Country	Tweets	Users
2018-10	Russia	9,041,308	3,667
2019-01	Russia	920,761	361
2019-01	Bangladesh	26,212	11
2019-01	Iran	4,671,959	2,496
2019-01	Venezuela	8,950,562	987
2019-04	Ecquador	700,240	787
2019-04	Saudi Arabia	340	6
2019-04	Spain	56,712	216
2019-04	UAE	1,540,428	3,898
2019-04	Venezuela	1,554,435	611
2019-06	Russia	2,288	3
2019-06	Iran	4,289,439	4,238
2019-06	Catalonia	10,423	77
2019-08	China	3,606,186	890
2019-09	China	10,241,545	4324

their ongoing information and propaganda operations, the social media platforms themselves have been forced to take action.

One of the actions that Twitter took is to suspend accounts associated with state sponsored propaganda campaigns and then release this data to the public for analysis and transparency. So far they have only released data associated with state sponsored manipulation and not other actor types. A summary of the data that they released is provided in Table 1 below. The largest and most prominent of these is the data associated with the Russian Internet Research Agency (IRA) and the Chinese data. The IRA data includes a well documented information campaign to influence an election and otherwise cause division in the United States, and the Chinese data is associated with information manipulation around the Hong Kong protests.

Our analysis of the Chinese and IRA data is a means for us to begin developing repeatable ways to characterize malicious online actors. Our experience is that social cybersecurity analysts often use a supervised machine learning algorithm to conduct their initial triage of a specific social media stream, say a stream related to an election event. This supervised model will often label tens of thousands of accounts as likely automated/malicious, which is still too many to sift through manually. While there are ways for an analyst to prioritize this list (for example finding the intersection of the set of likely bots with the set of influential actors measured with eigenvector centrality), it would be nice to characterize these malicious actors in a richer way than binary classification of “bot” or “not”. This paper, using the IRA and Chinese data to illustrate, will pave the way for future research and tools that will provide a comprehensive bot-labeling workflow for characterizing malicious online actors.

The IRA data that we will study in this paper is the original data set that Twitter released under their then nascent elections transparency effort. This release was

spurred by the fall-out after the 2016 US election and increasing evidence of Russian manipulation. The data has been studied as part of the Mueller Special Counsel investigation as well as several independent analysis conducted on behalf of the US Senate.

The Chinese data was produced from behind China's firewall and based on the IP addresses associated with the activity Twitter believes was produced by the People's Republic of China or a sanctioned proxy. This manipulation was attempting to change the narrative of the Hong Kong protest both for the residents of Hong Kong as well as the broader international community.

Before we spend some time going into a deeper comparison of these two data sets, we acknowledge that at a macro level they are very different because the target events are vastly different. In the case of the Russian IRA data, they were attempting to create a change in a foreign election on the other side of the world. In the Chinese case, they were largely trying to control the narrative of domestic events evolving inside their own borders. Acknowledging this macro level difference will shed some light on the other differences we uncover in this paper.

In addition to analyzing the *core* data that Twitter released to the public, we also collected additional data on all accounts that are mentioned, retweeted, replied to, or otherwise associated with the *core* data. This additional data was collected with the Twitter REST API, and throughout this paper we will refer to it as the *periphery* data. Note that this *periphery* data includes both malicious and non-malicious accounts. The malicious accounts have not been suspended by Twitter, and are either continuing to conduct information warfare or are in a dormant state waiting to be activated. The non-malicious accounts are accounts that became associated with the *core* data through a mention, retweet, or reply. These are often online actors that are either amplified or attacked in the information operation, or they could be innocent bystanders that bots and trolls mention in an attempt to build a following link so that they can influence them. Note that at the end of this paper we will attempt to estimate the number of accounts in the *periphery* data that are malicious and still active.

While several papers and reports as well as news articles have explored each of these data sets individually, as of the time of this writing we have not found a paper or report that expressly compares them. In conducting this research, our goal in order of priority is to:

1. Develop repeatable workflows to characterize information operations
2. Compare and contrast Russian and Chinese approaches to influence and manipulation of Twitter
3. Build on existing analysis of these unique data sets and the events and manipulation they are associated with

In order to characterize and then compare and contrast these data sets, we will develop and illustrate the use of social cybersecurity analytics and visualization. In this paper we will specifically focus on visual network analysis, new geographic analysis using flag emojis, temporal analysis of language and hashtag market share, bot analysis using several supervised machine learning models, meme

analysis of image memes, and analysis of state sponsored media involvement. We will then finish up by analyzing and discussing the number of accounts in the periphery data that are still conducting or supporting state sponsored information manipulation. Research such as this is key for threat assessment in the field of social cybersecurity [19].

## 2 Literature Review

Several reports and research papers have explored the data that Twitter released relative to the Russian/IRA and Chinese information operations. These are discussed below.

### 2.1 *Russia Internet Research Agency Data*

Russia's Internet Research Agency (IRA) is a St. Petersburg based company that conducts information operations on social media on behalf of the Russian government and businesses. The company began operations in 2013 and has trained and employed over 1000 people [12].

The IRA data has had more time and research effort than the newer Chinese manipulation data. Even before Twitter released the data to the public they allowed several research organizations an early analysis to accompany the release. Notable among these preliminary and largely exploratory analysis is the research by the Digital Forensic Labs [17].

The Special Investigation "Mueller" report, released on April 18, 2019, detailed the IRA operations [16]. The 443 page report contains 16 pages dedicated to IRA manipulation of information surrounding the 2016 US Presidential election. The manipulation detailed in the redacted report includes organization of grassroots political efforts, use of accounts masquerading as grass roots political efforts. The report indicates that the IRA accounts posed as anti-immigration groups, Tea Party activists, Black lives matter activists, LGBTQ groups, religious groups (evangelical or Muslim groups), as well as other political activists. It also detailed the methods used and organization of personnel against these methods. Two IRA employees received visas and traveled to the United States in order to better understand the social, cultural, and political cultures. IRA employees operated accounts initially focused on Twitter, Facebook, and Youtube but eventually including Tumblr and Instagram accounts. It also details the purchase of advertisements. It details a separate bot network that amplified IRA inauthentic user content. It noted that celebrities, politicians, and news outlets quoted, retweeted, or otherwise spread IRA messaging. The report outlines throughout the 16 pages how messaging for Trump was positive and supportive while the messaging for Clinton was negative. The IRA

was also central to the February 2018 indictment of 13 Russian nationals by Special Counsel Robert Mueller [1].

The second report regarding the IRA was conducted by New Knowledge at the request of the US Senate Select Committee on intelligence (SSCI) and focused on Facebook, Instagram, Twitter, Youtube, Google+, Gmail, and Google Voice involving the IRA. The report also shows some evidence of IRA activity on Vine, Gab, Meetup, VKontakte, and LiveJournal. The data that Twitter provided to New Knowledge was roughly the same data that was released to the public, but was not hashed and contained IP address and other information. This highlights the IRA switch from Facebook/Twitter to Instagram following their negative publicity. It highlighted that Instagram outperformed Facebook, highlighting the importance of images and memes in information operations. Like the Mueller report it highlights targeted communities. It also discusses voter suppression operations, such as encouraging voters to vote for a third candidate, stay home on election day, or false advertisements for voting on Twitter. In addition to highlighting pro-Trump and anti-Clinton campaigns, it also highlights activity meant to divide, such as secessionist messaging. It then conducts temporal analysis, URL analysis, and other content analysis. They highlight some of the tactics, branding, and recruitment. It also highlights the IRA's attacks against Republican primary candidates. They conduct extensive analysis of the memetic warfare. They highlight the IRA tactic of amplifying conspiracy theories. Finally, they thoroughly highlight efforts to divide America through secession ("if Brexit, why not Texit"). To summarize their analysis was primarily content, strategy, and effects across a sophisticated campaign that targeted Black, Left, and Right leaning groups [12].

The Computational Propaganda Project, like New Knowledge, was provided data by the US Senate Select Committee on Intelligence, to include the Twitter IRA data. In addition to temporal analysis, categorical analysis, target population identification, limited network analysis, hashtag and content analysis, It focused on cross platform activity [14].

Several other notable research efforts on the IRA include Arian Chen's lengthy New York Times Article entitled "The Agency" which details how the IRA organizes false alarms such as their Columbian Chemicals Explosion Hoax and the Ebola virus hoax [9]. Badawy et al conducts research of the 2016 IRA data and analyzes to what extent the effort supported the political left versus the political right [2], and is probably the closest article to the effort that we propose. Note that the Badawy effort only focuses on IRA data, and does not include any discussion of the Chinese data.

## ***2.2 Chinese Manipulation of Hong Kong Narrative***

In August 2019 Twitter released data associated with information and platform manipulation by the Chinese government around the Hong Kong protests. Twitter claims this was a state-backed information operation. As evidence for this claim,

they point to the fact that all of the activity and the associated IP addresses on the suspended accounts originated from within the People’s Republic of China (PRC) even though Twitter is blocked by the PRC (i.e. China’s ‘Great Firewall’). While some users in China access Twitter through VPNs, the nature of VPNs means the IP addresses aren’t from within the PRC. Twitter suspended the accounts for violating terms of service [18]. Censorship, while well documented, is difficult to measure [13].

The China data has had limited reporting on it. This is partially because it is newer, and also because it is harder to put together a cohesive picture of the data. Any cursory exploratory data analysis will often leave the researcher puzzled. Multiple posts on social media and elsewhere express this puzzlement. This is because the highest languages in the data are Indonesian, Arabic, and English, not Chinese. The most common hashtag is PTL (“Praise the Lord”). A substantial part of the data appears to involve an escort service or prostitution ring in Las Vegas, Asia and possibly elsewhere. It is only after extensive analysis that we will walk through in this report that the true nature of the data becomes evident.

While there are limited reporting on this data, we do want to call attention to the most thorough analysis we’ve found to date. The most comprehensive analysis we’ve found was conducted by Uren et al at the Australian Strategic Policy Institute [22]. This research highlights that these accounts attacked political opponents of the Communist Party of China (CPC) even before they began influencing the events in Hong Kong. Some of the primary conclusions of the report is that the Chinese approach appears reactionary and somewhat haphazard. They did not embed in virtual groups and slowly build influence, but rather generated simple spam that supported their messaging. This report does go into extensive temporal and geographic analysis that we will at times enhance but not duplicate. They do highlight that the lack of sophistication may be because it was outsourced to a contractor or because the government agency overseeing the operation lacked a full understanding of information operations. This report also highlights the fact that many of these accounts appear to be purchased at some point in their history. The authors show that 630 tweets contain phrases like ‘test new owner’, ‘test’, ‘new own’, etc. which are commonly used to show that a given account has come under new ownership.

### 3 Data

Twitter is a core platform for the global conversation, providing an open market for opinions and beliefs. By 2014 Twitter surpassed Facebook citations in the New York Times and by 2016 the New York Times cited Twitter more than twice as much as Facebook [23]. Online media often include Twitter posts of celebrities, politicians, and other elites in their content. To some extent, Twitter captures more of the global conversation (particularly in the West) while Facebook captures more of the local and topical conversations. Given this important opinion market, numerous

**Table 2** Summary of data

	IRA		China	
	Core	Periphery	Core	Periphery
Tweets	9,041,308	47,741,450	3,606,186	32,616,654
Users	3,667	667,455	890	20,4145
Top 5 languages	ru,en,de,uk,bg	en,ru,es,de,ar	in,ar,en,pt,zh	en,ar,pt,in,es

actors attempt to market their ideas and at times manipulate the marketplace for their benefit.

As mentioned above, the data is divided into the *core* data that Twitter released, as well as the *periphery* data that was associated with the *core* data. The *periphery* data includes any account that was mentioned, replied to, or retweeted by the *core* data. For every account in the periphery data, we collected the associated timeline (up to last 200 tweets). A summary of the *core* and *periphery* data sets is provided in Table 2.

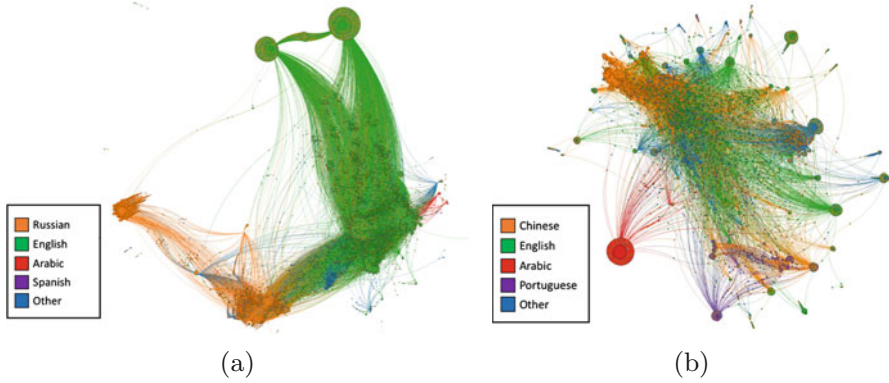
## 4 Characterization and Comparison

### 4.1 Network

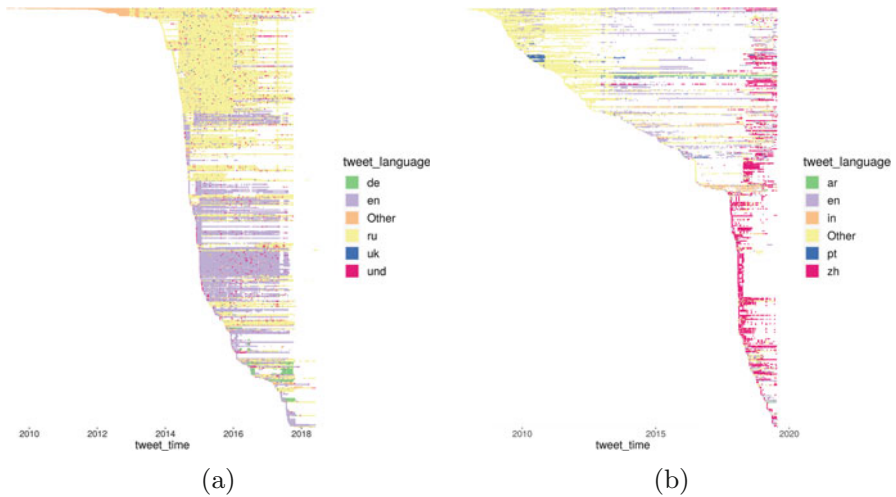
Information operations by their very nature manipulate narratives and networks. In order to understand and characterize them, we must understand the network that they are embedded in. To do this, we used the core data that was suspended and released by Twitter, and developed a network of communications. Links in the network represent one of the directed communication actions a user can take on Twitter, namely mention, reply, and retweet. These are communication links, not friend/following links. Nodes in this graph include both *core* and *periphery* accounts.

The networks are seen in Fig. 1a, with nodes colored by their most recent language. In the case of the Russian IRA data, we see clear lines of effort in Russian and English. When we zoom in on some of the Russian language clusters, we observe cascade communications that appear to be algorithmically created.

The conversation in the accounts used by the Chinese information operations is more complex primarily due to the fact that these accounts seem to be recently purchased by the Chinese government or government proxy, and the earlier histories of these accounts is varied. We observe that, even though Arabic and Portuguese have a large proportion of the conversation by volume, their use is relegated to a few accounts that are structurally segregated from the rest of the network. The Chinese and English language campaigns are much more intertwined as China directs their information campaign at the Western world and at Hong Kong. While the messaging is aimed at Hong Kong, it is not necessarily aimed internal to China since Twitter is blocked by China's firewall.



**Fig. 1** The conversational network of the core accounts suspended and released by Twitter (colored by most recent language used by account). (a) Russia core conversation (b) China core conversation



**Fig. 2** Tweets over time colored by language. (a) Russia. (b) China

## 4.2 History of Accounts

In this section we will detail the history of these accounts. We believe that Fig. 2 produces a good backdrop to explaining each of these campaigns and the differences between them. Each row in this graph is a single account with its tweets represented as points over time. This is colored by language (top 5 languages).

In the case of the Russian IRA, the timeline demonstrates a persistent effort to embed in both Russian and English language societies. Specific accounts embedded into target cultures and subcultures, learned to interact within the values and beliefs of the subculture, and then began to manipulate both the narrative and the network




in these subcultures. We do see some evidence of dormancy with some accounts leaving the conversation for sometimes years at a time, but nonetheless consistent effort to curate virtual persona's within a narrow context.

In the case of the Chinese disinformation effort, we see a very different approach. These accounts use multiple languages, exhibiting that these personas are not consistently embedding in the same networks and conversations. We also see long dormancy periods where these accounts are likely waiting to be activated or sold to a new bot handler. Then suddenly they all appear to be acquired or otherwise activated and begin tweeting in Chinese. This narrative accounts for the wide variety of languages and topics that baffled the cursory data explorer.





The history of the accounts shows a very different approach between the two information campaigns. The Russian effort demonstrates a planned and persistent effort to embed into the target society, and especially within target subcultures. They did this in the Russian language to manipulate their own population, and in English to manipulate beliefs and actions in America. Once embedded these agents continued to develop a following and influence a larger and larger swath of the American populace.

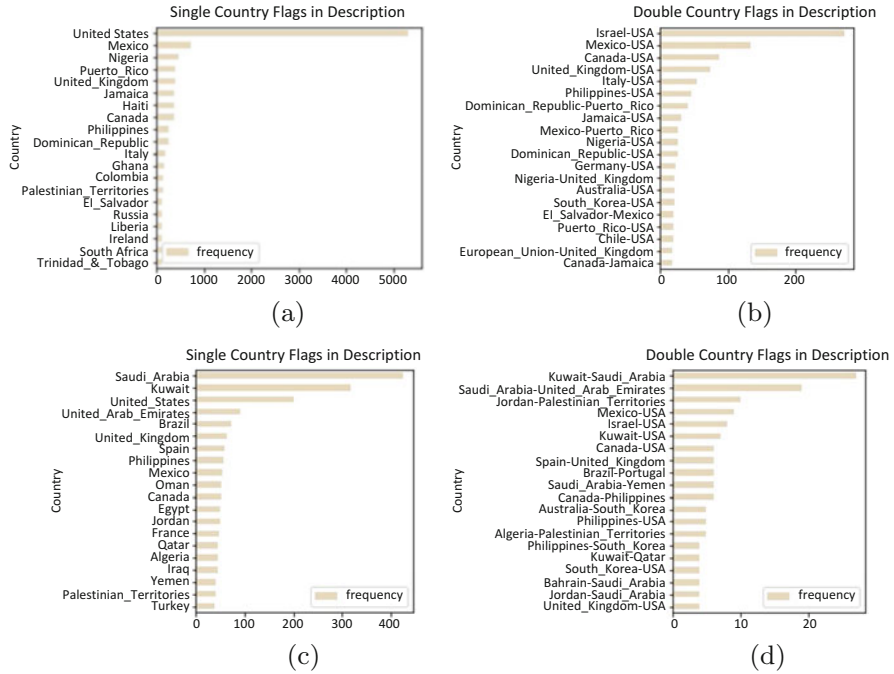
The Chinese approach was much more reactionary, seems less planned, and did not have any persistent effort to embed in networks to affect influence. To some extent, the Chinese effort was simply to spam their narrative across the international community.

### ***4.3 Geography of Accounts***

While the geography of both of these data sets have been explored to some extent, we wanted to take a little different approach to the geography of Twitter data. In our analysis here, we focus on the national flags that are often added to an actor's description field in Twitter. These flag emoji's are produced by using ISO 3166-1 internationally recognized two-letter country codes. Examples of flag emoji's are shown here . Flags are naturally used by individuals to associate themselves with a national identity. At times, individuals use multiple national flags in their description. Multiple national identities may be the result of immigration or a proud ex-patriot.

In our analysis of disinformation streams, however, we've seen bots and other malicious accounts use two or more flags in their profile. We believe that this is done so that an actor can leverage a curated and popular account in multiple target audiences and conversations. In particular we've seen this done with accounts so that they can participate in political conversations in North America and Europe, possibly in different languages, and make it look as if they're just a passionate ex-patriot.

We found evidence of this behavior in the core data set, particularly in the IRA data. Two examples are  in  and Russian  living in the US . In these cases, a description like this allows the casual observer to rationalize why the account



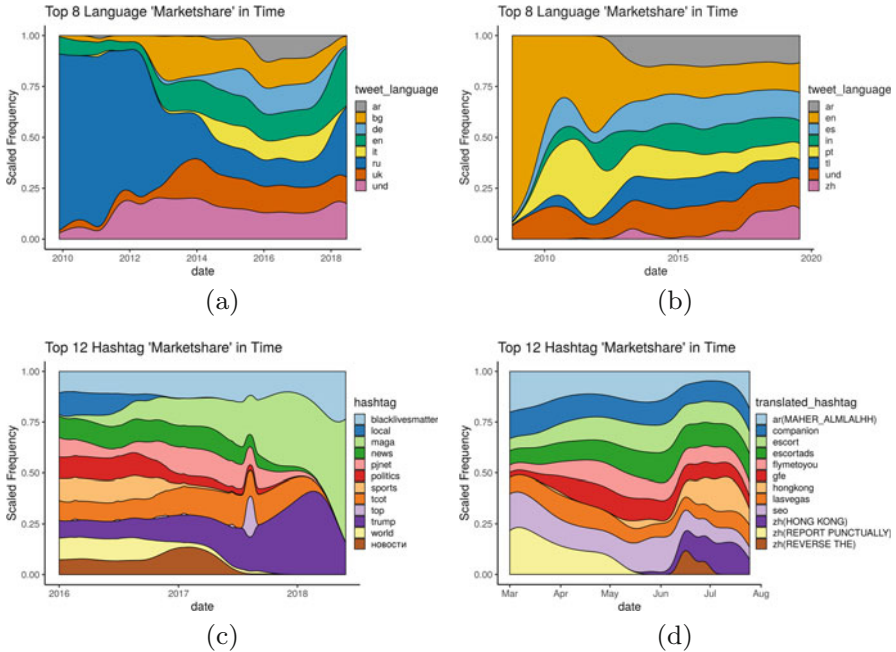
**Fig. 3** The distribution of flag Emoji’s in account descriptions. The high volume of unexpected flags used for the China data (such as Kuwait/Saudi Arabia) is due to the fact that many of these accounts were recently purchased by the Chinese government, and therefore most tweets and account descriptions by these accounts were produced by their previous owners. **(a)** IRA single flag. **(b)** IRA double flag. **(c)** China single flag. **(d)** China double flag

switches back and forth between Russian and English and between Russian social/political conversations and American social/political issues.

To explore this at scale, we developed algorithms that would extract the flag emoji’s and build distributions. When we did this we built a distribution of single occurring flags and then of multiple flag combinations. The results of this analysis are provided in Fig. 3. In particular we see a high number of US-Israel flags combinations among the Russian information operations. Also of note is a high number of US-Italian combinations. While many of these may be legitimate, we have observed some accounts in different data that are simultaneous meddling in US political debate in English while encouraging Italy to leave the European Union in Italian.

#### 4.4 Calculating Content Marketshare Over Time

Although we’ve already looked at the histories of these accounts, we wanted to understand temporal distributions better so that we can understand how these



**Fig. 4** Normalized marketshare of language and hashtags for core IRA and Chinese data suspended by Twitter. (a) IRA language market share. (b) China language market share. (c) IRA hashtag market share. (d) China hashtag market share

accounts were used over their life span as well as in the world events they’re respectively associated with. To do this we explored the use of language and content over time with temporal market share.

To compute the temporal market share of language and hashtags we identified the top 8 languages and the top 12 hashtags in the core data for each operation, and their normalized portion (or market share) of the conversation over time. We see the visualization in Fig. 4. In the IRA data (graphs on left), we see a clear transition of information operations conducted in Russian to begin manipulation in Ukraine, English and other languages almost exclusively focused on Europe and the West.

In the plot of IRA hashtag market share, two things jump out. The first is the sudden outsized growth of IRA support of the #MAGA hashtag and the American right. The IRA did infiltrate the American left, but not to the same extent as the American right. The second and equally alarming observation is the long term and persistent use of the #blacklivesmatter hashtag as some of the IRA agents embedded into the African American subculture. The final but equally important observation we see here is that many of the hashtags are associated with a standard news organization. Multiple accounts in the data attempted to appear as a local news source or news aggregator in order to have the appearance of legitimacy.

From the Chinese core data, we see a wide variety of languages with only a small uptick in Chinese language at the end. Likewise the hashtag plot only has a small uptick in English and Chinese use of Hong Kong at the end. While Twitter associated all of the accounts with deliberate operations by the Chinese, the actual volume of data associated with the Hong Kong protests is limited compared to the total volume over the life of these accounts.

## 4.5 Bot Analysis

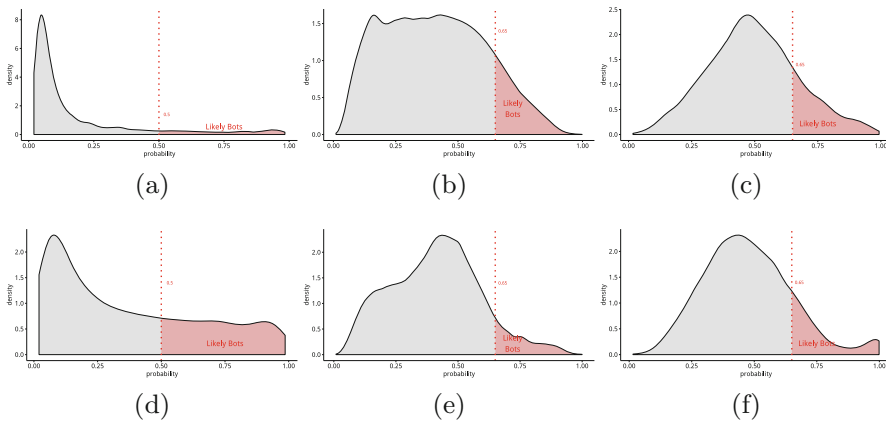
Social media bots are any account that has some level of action being automated by a computer. On Twitter tweeting, retweeting, replying, quoting, liking, following, and general searching can all be automated. In this section we leverage several bot detection tools to predict the number of accounts that appear to have automated behavior. Memes and bots are tools used to conduct information maneuvers in influence campaigns [6].

The models used below are two external models as well as two that were developed by our team. The first external model is the Debot model [8]. The Debot model is an unsupervised model that finds bots that are correlated using warped correlation. In other words, this model finds two or more accounts that are posting the same content at roughly the same time. The Debot team continually monitors parts of Twitter, and keeps a database of accounts that they've found to be correlated. In our search through the Russia and Twitter periphery data, we searched the Debot database to identify any of our accounts that have been found before. The second external model is the Botometer model (previously called the BotOrNot model) [10]. The Botometer model is a supervised machine learning model with well documented feature space. The Botometer Application Programming Interface (API) accepts a user ID or screen names as input, scrapes the Twitter API using the consumer provided keys on the server side, and then returns a score for content, friends, network, sentiment, temporal, user, and the universal score for the account. Given this method, Botometer scores are only available for accounts that are still active (i.e. not suspended, private, or otherwise shutdown). Due to the time required to scrape the timeline, in both of our data sets we randomly sampled 5,000 accounts for the Botometer model.

We've also listed scores for two models developed internally. The Bot-Hunter suite of tools provides supervised bot detection at several data granularities. Tier 1 conducts bot detection with a feature space developed from the basic tweet JSON data that is returned by the Twitter API [4]. This includes features extracted from the *user* object and the *tweet* object. Tier 2 performs bot detection using the users timeline (adding more content and temporal features), and Tier 3 uses the entire conversation around an account to predict the bot score [3]. Due to the computational cost of running Tier 3 (approximately 5 min per account), it is best for only a handful of accounts and was not used on these data sets. The Bot-Hunter Tier 1 models was run on all data, and the Tier2 was run on a random sample of

**Table 3** Bot prediction for *core* and *periphery* data (% of total)

	Russia IRA		China (Hong Kong)	
	Core	Periphery	Core	Periphery
Accounts		697,296		204,920
Debot	**	1.07%	**	0.66%
Botometer	**	9.1 ± 0.7%	**	28.5 ± 1.3%
Bot-Hunter Tier 1		13.20%		8.68%
Bot-Hunter Tier 2	9.35%	15.9 ± 0.9 %		13.8 ± 0.9%
Suspended/Closed	100%	4.30%	100%	0.30%



**Fig. 5** Probability distributions for bot prediction for Botometer, Bot-Hunter(BH) Tier 1 and Tier 2 with threshold shown. (a) IRA botometer. (b) IRA BH-tier 1. (c) IRA BH-tier2. (d) China botometer. (e) China BH-tier 1. (f) China BH-tier 2

5000 accounts. Note that unlike Botometer, Bot-Hunter runs on existing data and was therefore able to predict on core, periphery, and suspended accounts. We’ve also developed an abridged version of Bot-Hunter Tier 1 that can run on the core data since it doesn’t contain all features available for the unabridged model.

From Table 3 we see that models predict that 9–15% of the Russian core and periphery have likely automated behavior, with Hong Kong estimates slightly lower with Bot-Hunter predicting 8–14% automated behavior and Botometer as the outlier with 28% prediction.

We get even more insight into these models and data by looking at Fig. 5. This shows the probability distribution and chosen thresholds for each of the models on the periphery data. The biggest takeaway in these images is the difference between the shape of the Botometer model and the Bot-Hunter models. Although both are trained with a similar supervised learning model (Random Forest Classifier), they were trained on very different training data. Because of this, Botometer shows that most accounts are very unlike automated accounts, whereas Bothunter models show that the majority of accounts seem to appear a little more automated. Given that both

models are similar, these distributions are saying that the these suspect accounts associated with Russian and Chinese disinformation are more similar to the data that Bot-Hunter was trained on than the data that Botometer was trained on.

## 4.6 Multi-media Analysis

Richard Dawkins originally created the word meme in his book *Selfish Gene* in which he defined a meme as a "... noun that conveys the idea of a unit of cultural transmission, or a unit of imitation" [11]. Shifman later adapted and defined internet memes as artifacts that "(a) share common characteristics of content, form, and/or stance; (b) are created with awareness of each other; and (c) are circulated, imitated, and transformed via the internet by multiple users" [20, 21].

Internet memes, particularly multi-media memes, are increasingly used in online information warfare. This phenomena has been highlighted in articles like the *New York Time* "The Mainstreaming of Political Memes Online" [7], and has been dubbed memetic warfare. To analyze memes in these two data sets, we developed a deep learning meme classifier to extract memes from the multi-media archives that Twitter shared along with the data. We ran this classifier on all images in the IRA data set, and on all Hong Kong related images in the China data set. Examples of IRA memes are provided in Fig. 6 and examples of China memes are provided in Fig. 7.

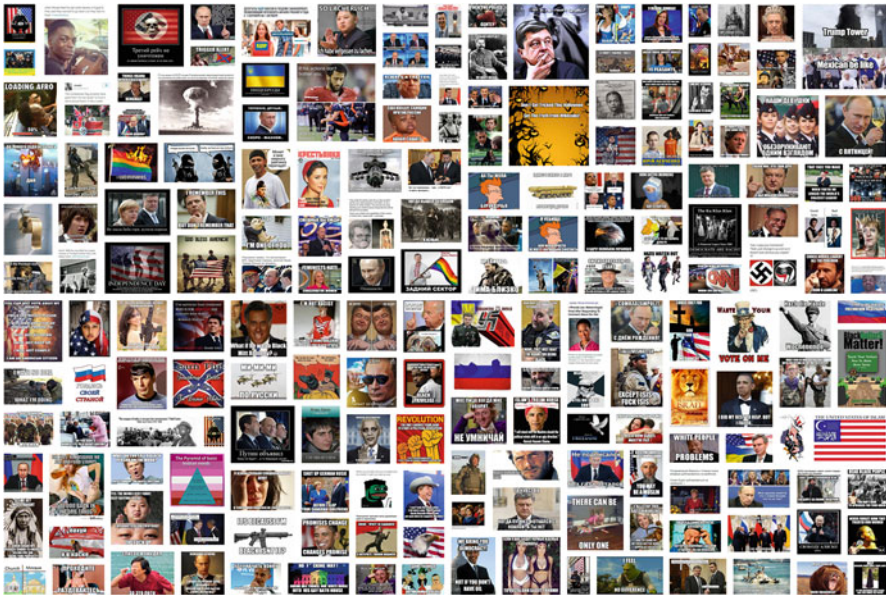


Fig. 6 Russian IRA memes





**Table 4** Add caption

State owned media	IRA core data	Chinese core data
Russian	72,846	8
Chinese	226	1,400
American	11	0
Korean	0	2
German	62	2

worldwide penetration of RT, Sputnik, and other state sponsored news agencies, while China has been gaining greater international penetration with China Xinhua News. To measure the extent that to which this data is amplifying these voices, we collected a large list of all Twitter handles associated with these Russian and Chinese state owned media companies, as well as handles associated with several other country’s state owned media (for example the US Voice of America) for comparison. While the degree to which each of these handles spread state “propaganda” varies widely, we provide them for comparison.

We then scanned the *core* data for both datasets to examine the degree to which each data set amplifies these state owned media company’s. The results are provided in Table 4.

As can be seen by this table, both the Chinese and especially the Russian dataset provides massive amplification for these state owned media.

## 5 How Many Similar Actors are Left?

One of the biggest questions that remains after going through this data is “How many state sponsored actors are still at large in the virtual world and currently manipulating world events?” To try to answer this question we spend some time analyzing the *periphery* data that is still mostly ‘alive’ and active on Twitter. Some of these actors may have been randomly brought into the data set, possibly by bots that were randomly mentioning normal citizens on Twitter in an effort to build a following/friend tie and begin to influence them. Others, however, are undoubtedly part of the larger information campaign and are still conducting malicious and divisive operations.

As shown above, at ~10% of both streams exhibit bot like behavior (these are again conservative estimates). Of the accounts in the periphery, 85.2% of the Russian accounts and 64% of the Chinese accounts are active, meaning they are not dormant and have tweeted in the last 6 months. Additionally, these accounts continue to amplify state owned propaganda. The IRA *periphery* amplifies Russian state owned media 6,023 times, and the China *periphery* amplifies Chinese state owned media 1,641 times.

Below we try to capture the primary topics that these accounts are embedding in. To do this we sampled 5000 accounts from the periphery of Russia and from China,





**Fig. 8** Current Information Operations by Russia found in the *periphery* data. (a) Possible Russia influence in the US Left (8–10% of Accounts). (b) Possible Russia influence in the US Right (10–12% of Accounts)

collected the last 200 tweets associated with each accounts. After selecting only those tweets in the last 6 monthsh, we conducted topic analysis with latent dirichlet allocation (LDA). By optimizing the Calinski Harabaz score, we chose a  $k$  of 10 for LDA.

The Russian data shows clear topic groups that are attempting to meddle in Western affairs. The wordclouds of two of these topic groups is shown in Fig. 8. These images show a continued effort to divide America by further polarizing an already polarized political climate. Note that other topics not shown here include efforts to meddle in Europe (particularly amplifying the voice of the Yellow Vest Movement as well as far right groups), meddle in Canadian elections (clearly seen in the prominent place of #cdnpoli and #elxn43 in one LDA topic group of every sample tested).

From this we find that the Chinese data is still too diverse. The *periphery* data is associated with the entire timeline of these accounts, and is therefore too diverse to define clear information operation efforts and identify them in topics. During LDA and further analysis we found  $\sim 190$  K accounts associated with Hong Kong, but they seemed to be across the spectrum of the discussion without any strongly coordinated disinformation operations (at least not in this *periphery* data). With the LDA analysis, we did find one sizable group that appeared to be against the current US administration. Once again, because of the randomness of the data it was difficult to claim this was due to a coordinated effort and not just caused by random bot behavior.

## 6 Conclusion

Throughout the data we see an experienced, sophisticated and well resourced campaign by Russia’s Internet Research Agency while we also observe a Chinese campaign that appears reactionary and ad hoc. Several major conclusions are summarized below:

- The IRA's effort included identification and study of target subcultures with significant effort to shape messaging to leverage existing biases.
- The Chinese effort was aimed at Hong Kong and the international community at large without evidence of extensive effort to identify or a target audience or craft messaging for a specific audience.
- The IRA effort demonstrates an understanding of internet memes and a willingness to take risks in releasing multi-media messaging that will evolve in the masses.
- The Chinese effort demonstrates an unwillingness to release internet memes that will evolve outside of the direct control of central authorities.
- Both efforts, but particularly the Russian effort, demonstrate an effort to use these covert information operations to enhance the overt information operations conducted by state owned media companies.

While the focus of this research is on manipulation by well resourced nation-states, these same tactics can and are being used by smaller nation states (Saudi Arabia, Iran, Venezuela) and by non-state actors such as ISIS.

This work lays the foundation for building a repeatable end-to-end process for characterizing malicious actors in disinformation streams in social media, which is essential for national security [5]. These efforts to characterize actors will assist social cybersecurity analysts and researchers in getting beyond the binary classification of 'bot or not.' Future research will describe and illustrate this full workflow and several different data sets.

**Acknowledgements** This work was supported in part by the Office of Naval Research (ONR) Award N00014182106 and Award N000141812108, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

## References

1. Alvarez, P., Hosking, T.: The full text of Mueller's indictment of 13 Russians'. *The Atlantic*, 16th Feb (2018)
2. Badawy, A., Addawood, A., Lerman, K., Ferrara, E.: Characterizing the 2016 Russian IRA influence campaign. *Soc. Netw. Anal. Min.* **9**(1), 31 (2019)
3. Beskow, D., Carley, K.M.: Bot conversations are different: leveraging network metrics for bot detection in twitter. In: *Advances in Social Networks Analysis and Mining (ASONAM)*, 2018 International Conference on, pp. 176–183. IEEE (2018)
4. Beskow, D., Carley, K.M.: Introducing bothunter: a tiered approach to detection and characterizing automated activity on twitter. In: Bisgin, H., Hyder, A., Dancy, C., Thomson, R. (eds.) *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer (2018)
5. Beskow, D.M., Carley, K.M.: Army must regain initiative in social cyberwar. *Army Mag.* **69**(8), 24–28 (2019)

6. Beskow, D.M., Carley, K.M.: Social cybersecurity: an emerging national security requirement. *Mil. Rev.* **99**(2), 117 (2019)
7. Bowles, N.: The mainstreaming of political memes online. *New York Times* (Feb 2018). <https://www.nytimes.com/interactive/2018/02/09/technology/political-memes-go-mainstream.html>
8. Chavoshi, N., Hamooni, H., Mueen, A.: Debot: twitter bot detection via warped correlation. In: *ICDM*, pp. 817–822 (2016)
9. Chen, A.: The agency. *N. Y. Times* **2**(6), 2015 (2015)
10. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botnotot: a system to evaluate social bots. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 273–274. International World Wide Web Conferences Steering Committee (2016)
11. Dawkins, R.: *The Selfish Gene: With a New Introduction by the Author*. Oxford University Press, Oxford (2006). (Originally published in 1976)
12. DiResta, R., Shaffer, K., Ruppel, B., Sullivan, D., Matney, R., Fox, R., Albright, J., Johnson, B.: *The Tactics & Tropes of the Internet Research Agency*. New Knowledge, New York (2018)
13. Faris, R., Villeneuve, N.: Measuring global internet filtering. In: *Access Denied: The Practice and Policy of Global Internet Filtering*, vol. 5. MIT Press, Cambridge (2008)
14. Howard, P.N., Ganesh, B., Liotsiou, D., Kelly, J., François, C.: *The IRA, Social Media and Political Polarization in the United States, 2012–2018*. University of Oxford, Oxford (2018)
15. McDonnell, S.: Why China censors banned Winnie the pooh – BBC news. <https://www.bbc.com/news/blogs-china-blog-40627855> (July 2017). Accessed 29 Sep 2019
16. Mueller, R.S.: Report on the Investigation into Russian Interference in the 2016 Presidential Election. US Department of Justice, Washington (2019)
17. Nimmo, B., Brookie, G., Karan, K.: #trolltracker: twitter troll farm archives – DFRLAB – medium. [file:///Users/dbeskow/Dropbox/CMU/bot\\_labels/references/ira/%23TrollTracker\\_%20Twitter%20Troll%20Farm%20Archives%20-%20DFRLab%20-%20Medium.html](file:///Users/dbeskow/Dropbox/CMU/bot_labels/references/ira/%23TrollTracker_%20Twitter%20Troll%20Farm%20Archives%20-%20DFRLab%20-%20Medium.html). Accessed 23 Sep 2019
18. Safety, T.: Information operations directed at Hong Kong. [https://blog.twitter.com/en\\_us/topics/company/2019/information\\_operations\\_directed\\_at\\_Hong\\_Kong.html](https://blog.twitter.com/en_us/topics/company/2019/information_operations_directed_at_Hong_Kong.html) (August 2019). Accessed 26 Sep 2019
19. National Academies of Sciences, Engineering, and Medicine: A Decadal Survey of the Social and Behavioral Sciences: A Research Agenda for Advancing Intelligence Analysis. The National Academies Press, Washington (2019). <https://doi.org/10.17226/25335>, <https://www.nap.edu/catalog/25335/a-decadal-survey-of-the-social-and-behavioral-sciences-a>
20. Shifman, L.: The cultural logic of photo-based meme genres. *J. Vis. Cult.* **13**(3), 340–358 (2014)
21. Shifman, L.: *Memes in digital culture*. MIT Press, London (2014)
22. Uren, T., Thomas, E., Wallis, J.: Tweeting Through the Great Firewall: Preliminary Analysis of PRC-linked Information Operations on the Hong Kong Portest. Australia Strategic Policy Institute: International Cyber Policy Center, Barton (2019)
23. Von Nordheim, G., Boczek, K., Koppers, L.: Sourcing the sources: An analysis of the use of twitter and facebook as a journalistic source over 10 years in the *New York times*, the *guardian*, and *süddeutsche zeitung*. *Digit. Journal.* **6**(7), 807–828 (2018)