Computational and Mathematical Organization Theory (2020) 26:465–483
https://doi.org/10.1007/s10588-019-09298-1

S.I.: SOCIAL CYBER-SECURITY

Interoperable pipelines for social cyber-security: assessing Twitter information operations during NATO Trident Juncture 2018

Joshua Uyheng[1] · **Thomas Magelinski**[1] · **Ramon Villa-Cox**[1] · **Christine Sowa**[1] · **Kathleen M. Carley**[1]

Published online: 28 September 2019

Abstract
Social cyber-security is an emergent field defining a multidisciplinary and multi-methodological approach to studying and preserving the free and open exchange of information online. This work contributes to burgeoning scholarship in this field by advocating the use of interoperable pipelines of computational tools. We demonstrate the utility of such a pipeline in a case study of Twitter information operations during the NATO Trident Juncture Exercises in 2018. By integratively utilizing tools from machine learning, natural language processing, and dynamic network analysis, we uncover significant bot activity aiming to discredit NATO targeted to key allied nations. We further show how to extend such analysis through drill-down procedures on individual influencers and influential subnetworks. We reflect on the value of interoperable pipelines for accumulating and triangulating insights that enable social cyber-security analysts to draw relevant insights across various scales of granularity.

Keywords Social cyber-security · Information operations · Interoperability

1 Introduction

Online social networks have allowed people to exchange information and express their views at an unprecedented scale. But while some scholars have argued that such platforms democratize public discourse, recent years have shown how adversarial actors may employ diverse strategies to manipulate public opinion toward disruptive social and political outcomes. Information operations are not a new phenomenon.

✉ Joshua Uyheng
juyheng@andrew.cmu.edu

[1] Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research, Carnegie Mellon University, Wean Hall, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

However, their adaptation to online settings uniquely capable of influencing large collectives has made them a more urgent area of concern. Such activities have been documented across high-stakes contexts such as national elections, potentially even posing a threat to international security.

How might researchers study information operations in a comprehensive and informative manner? Significant prior work has attempted to tackle this question from a variety of perspectives (Shu et al. 2017; Tucker et al. 2018; Zhou et al. 2019). While prevailing scholarship has primarily focused on developing state-of-the-art tools for solving social cyber-security problems one by one, scant work demonstrates how such tools may be integrated in practice. In the context of real-world online information operations, we argue that the utility of single tools is limited. In contrast, methodological pipelines emphasizing the interoperability of such tools may yield multifaceted insights better-suited to the complexity of their object of study (Conroy et al. 2015; Michelucci et al. 2015; Uyheng and Carley 2019). This paper therefore presents a framework for building interoperable pipelines of computational tools for analyzing online information operations. We turn to the Twitter conversation surrounding the NATO Trident Juncture Exercise in 2018 as an illustrative case study.

The succeeding sections of this paper are structured as follows. First, we review prior work in social cyber-security that illustrates the value of existing computational approaches to studying online information operations in Sect. 1.1. Next, we explain how cutting-edge techniques in machine learning, natural language processing, and dynamic network analysis can be utilized in a complementary fashion tailored toward different facets of information operations in Sect. 1.2. Following an overview of the NATO Trident Juncture Exercise in 1.3, we summarize how each component of our proposed pipeline maps to our Twitter dataset in Sect. 2. Finally, we present the results of our integrated analysis, assessing the activity of automated bot accounts, the messages they spread, and their influence in the overall public conversation in Sect. 3. With our findings, we concretely demonstrate how pipelined tools refine social cyber-security inquiry through accumulation and triangulation of multiple streams of analysis. We conclude with a discussion of advancing methodological frameworks for social cyber-security, as well as several practical considerations for working in concrete information operation settings.

1.1 Social cyber-security

Social cyber-security has been defined as a multidisciplinary and multimethodological field that studies how to preserve the internet as ‘a free and open space for the exchange of information’ (Carley et al. 2018). An important theme of research in this emergent field is the analysis of online content polluters, such as bots and trolls (Ferrara et al. 2016; Lee et al. 2011). Bots refer to automated online accounts while trolls are typically defined as human accounts which spread aggressive or disruptive messages (Beskow and Carley 2019a; Cheng et al. 2017). Both types of content polluters engage in a variety of strategies to achieve specific information operation objectives. Research in social cyber-security emphasizes the importance of

examining these activities through the lens of both computation (e.g., artificial intelligence, software engineering) and the social sciences (e.g., political science, social psychology), as both technological and societal issues are intertwined in determining the success and spread of such activities (Ferrara 2017; Lazer et al. 2018; Shu et al. 2017).

Prior research categorized these activities as either message-driven or network-driven (Beskow and Carley 2019b). On the one hand, message-driven activities focus on influencing public perceptions about various entities and events of interest. For instance, in electoral contexts, paid online accounts may be used to boost the profile of a favored candidate or spread rumors about an opponent (Bennett and Livingston 2018; Mejias and Vokuev 2017; Stewart et al. 2018). On the other hand, network-driven activities are interested in manipulating group dynamics to achieve certain purposes. Some operations may intend to isolate two groups in order to create polarized echo chambers (Garrett 2009; Nekmat and Lee 2018). In other cases, operations may work to encourage interactions between two groups so that their beliefs clash with each other in a phenomenon called trench warfare (Karlsen et al. 2017). By harnessing both message-driven and network-driven campaigns, information operations take advantage of the socio-technical landscape of online social networks to shape how information spreads and is perceived by online communities. Social cyber-security fundamentally aims to understand such dynamics to classify adversarial actors and their activities, assess and predict their impact, and design effective strategies for intervention and building the resilience of online communities.

1.2 Methods for analyzing information operations

A growing scholarship has tackled problems in social cyber-security with notable success. In this section, we briefly review this literature and identify methodological gaps that we propose can be resolved through the design of interoperable methodological pipelines (Wegner 1996). We go over how machine learning, natural language processing, and dynamic network analysis are used to detect information operations, characterize their content, and assess their overall influence and impact.

Bot detection: who are the bots? A basic issue in analyzing information operations deals with determining the number of bot accounts involved in an online conversation. Features of bot accounts can be difficult to define in concrete terms, but viewed in aggregate, patterns in their messaging behavior or relationships to other users in the social network can make them clearly discernible. Machine learning algorithms have been shown to be effective in elucidating such patterns automatically (Morstatter et al. 2016; Qi et al. 2018). Given a reasonably large dataset of labeled bot and non-bot accounts, predictive models can be trained to discriminate between each type of account with decent accuracy ($\geq 90\%$) across various contexts. Using random forest classifiers, Beskow and Carley (2019a) show how different features of Twitter accounts, from the use of random strings in their usernames to their network-level activities, can be incrementally used in a tiered approach for bot detection. Textual features have also been widely used to detect trolling, cyberbullying, and opinion manipulation on a variety of online contexts such as news community

forums and Twitter (Beskow and Carley 2018b, a; Mihaylov et al. 2015; Seah et al. 2015). Lee et al. (2010, 2011) further show not only how to jointly exploit user features, language features, and network features simultaneously, but also how to obtain more labeled data through the use of social honeypots. We draw upon off-the-shelf bot detection algorithms as a crucial step in our proposed interoperable pipeline for information operations analysis.

Topic modeling: what messages are they spreading? Another fundamental question researchers may pose is related to the messages bot accounts promote or disrupt. Latent Dirichlet allocation is a general algorithm for extracting topics from a corpus of texts (Blei et al. 2003). Topics are taken to be distributions over words, and a textual corpus is considered a mixture model over several such distributions. Techniques such as topic modeling offer a means of capturing core ideas being spread by bot accounts, thereby extracting the key aspects of information relevant to a disinformation campaign (Chew and Turnley 2017; Yang et al. 2015). Novel work has shown how discovering conflicts among modeled topics can be helpful for identifying false information (Jin et al. 2016).

Dynamic network analysis: what influence do they have? Finally, impact assessment of information operations may involve searching for influential participants in an online conversation and characterizing the messages they promote. Influencer analysis asks: Given a large-scale conversation on Twitter, which users impact the conversation the most? Dynamic network analysis offers a systematic framework for precisely quantifying influence in an online conversation. In Twitter, for instance, users are modeled as nodes of a graph. Depending on the behavior of interest, retweets between users may be taken as a directed edge between their corresponding nodes. By studying multiple types of nodes (e.g., bot accounts, news agencies) and edges (e.g., retweets, replies) simultaneously and over time, dynamic network analysis goes beyond a standard social network analysis framework to uncover more complex insights into online discourse (Carley et al. 2007). Under a network framework, measures of centrality may be used to determine important users (Riquelme and González-Cantergiani 2016). For example, a user can be labeled an influencer if they have a high in-degree in their following network, since their tweets reach many people. Users can also reach many people if they are retweeted by someone with a large following. Additional measures include but are not limited to the total number of retweets, favorites, or mentions a user receives (Bakshy et al. 2011; Dubois and Gaffney 2014).

Interoperable integration of computational tools In this paper, we propose that putting together different tools in an interoperable pipeline can generate comprehensive and relevant insights for social cyber-security. Prior work had shown how the complexities of information operations require multiple methods for comprehensive analysis. Arif et al. (2018), for instance, showed how graph-based identification of online communities can be augmented by interpretative analysis of discourse by disinformation agents to better understand the objectives of public opinion manipulation campaigns related to #BlackLivesMatter. Benigni et al. (2018) likewise showed how different types of dynamic network analysis, such as clustering of user networks or clustering of hashtag networks, can be used to identify different types of ideology and quantify levels of radicalization. Our framework goes further in showing

how a *combination* of computational tools—used in series as well as in parallel—can be used to drive human-driven analysis of information operations and holistically infer more complex insights than could be derived from single tools. We therefore showcase how interoperability uniquely benefits studies in the field of social cyber-security.

1.3 The NATO Trident Juncture Exercise 2018

The North Atlantic Treaty Organization (NATO) was founded in 1949 as an alliance of governments for the preservation of peace in the postwar era. The Trident Juncture Exercises (TRJE) are an opportunity for member nations to consolidate their forces in joint military exercises as a symbol of unity and commitment to their shared purpose. For the latest iteration of the exercises, from late October through November 2018, TRJE brought together 50,000 military and civilian personnel from 31 NATO and partner countries in Norway, making it NATO's largest exercise in two decades.

With its significant show of military force, TRJE attracted widespread discussion on mainstream and social media platforms. However, the same online attention also enabled the spread of manipulative messaging surrounding the exercises and NATO more generally. One long-standing issue encompasses NATO's tense relationship with Russia, especially in view of the exercise's geographic proximity to the superpower's borders. Another possible trigger for negative messaging, as covered extensively by the popular press, was the unforeseen crash of a Norwegian frigate, the Helge Ingstad, while returning from the exercise. Both Russia and the frigate collision represent sensitive concerns vulnerable to adversarial information operations online. Through the lens of social cyber-security, we study the NATO Trident Juncture as a prime example of an online event attracting contentious discourse and information operations, with high-level stakeholders, and international significance. Prior work on previous iterations of the Trident Juncture had also found evidence of influential information operations targeting the event on online platforms (Al-Khateeb et al. 2019; Carley and Beskow 2017).

2 Data and methods

2.1 Dataset

To assess the Twitter conversation surrounding NATO TRJE 2018, we analyzed 236,809 tweets collected from October 22, 2018 to November 13, 2018. Data collection was primarily conducted using the Twitter APIs with hashtags #tridentjuncture, #nato, and their non-English variants. Each tweet came with metadata on its corresponding user account and relevant interactions with other users and tweets (e.g., retweets, mentions). Hashtags for data collection were iteratively refined as additional hashtags of relevance were discovered from incoming data. However, we note that the Twitter API does not necessarily provide a purely random sample of the

entire Twitter conversation; thus, conclusions from this work should be considered in light of these standard limitations (Morstatter et al. 2013).

A total of 81,555 unique users were represented in the dataset. For certain time-sensitive aspects of the analysis, tweets were divided into four periods: phase 1, which included all tweets before the official start of TRJE on October 25; phase 2, which spanned the main days of the exercise from October 25 to November 7; phase 3, which encompassed the crash of the Helge Ingstad on November 8 and the two days afterward; and phase 4, which accounted for all tweets past November 10. We note that additional online conversations may have taken place after our cutoff dates; however, we heuristically finished data collection a week after the main exercises to concentrate on event-specific tweets. Future work may study more closely how such data collection decisions may adapt to the diffusion of information in relation to large-scale public events (Babcock et al. 2019).

2.2 Pipeline of tools

As visualized in Fig. 1, a series of interoperable tools was used to leverage textual, user, and interaction information for topic analysis, bot detection, role identification, location prediction, and influencer analysis.

Latent Dirichlet allocation Latent Dirichlet allocation (LDA) was performed to provide an exploratory characterization of the main topics of conversation and track their prevalence across the four time periods (Blei et al. 2003). Extensive text preprocessing was conducted to standardize contractions, remove URLs and stop-words, etc. The final number of topics was evaluated using the coherence score, which assesses the co-occurrence of words belonging to certain topics over others in similarly classified documents (Röder et al. 2015). By producing probabilistic scores assigning each tweet to a topic, as well as a list of keywords defining each topic, LDA outputs quantitatively provided a means to understand the general contents of online talk about NATO and TRJE. Manual review of keywords and tweets with the highest scores for each topic provided complementary, qualitative evidence for interpreting topics holistically (Montiel et al. 2019).

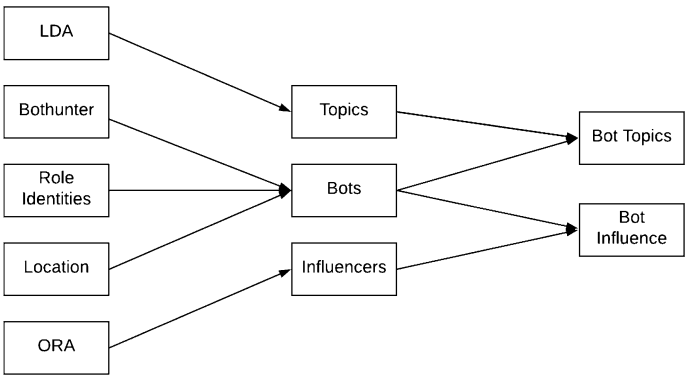

Fig. 1 Interoperable pipeline of tools for information operations analysis

Bothunter We used Bothunter, a random forest machine learning model, to examine Twitter accounts in the dataset and output a probability of the user being a bot based on their profile features (Beskow and Carley 2019a). Other work has shown that random forest models provide competitive performance across benchmark datasets (Varol et al. 2017), though we note that more complex models may provide slightly better performance for some contexts (Kudugunta and Ferrara 2018). As Bothunter produces a probabilistic (rather than binary) output, we select 60% as the threshold above which we decide a user is a bot. Heuristically, this is more strict than the standard 50% threshold in binary classification tasks; we choose a slightly higher threshold to increase precision (by reducing false positives), though potentially to the expense of recall (by increasing false negatives). We note that future work may explore altering this threshold to maximize performance alongside recall (Nazer et al. 2019). We also preferred Bothunter over other existing classifiers for its combined speed and accuracy.

Role identification Role identification employed a neural network model trained on a large dataset of user descriptions and tweets to classify accounts as belonging to special actor classes such as news agencies, reporters, government offices, and celebrities.

Location prediction Location prediction was also conducted using a neural network model trained on a dataset of user descriptions and known locations (Huang and Carley 2017).

ORA Finally, Organization Risk Analyzer (ORA) enabled the analysis of multimodal networks to identify influential users, as well as characterize the overall structure of the Twitter conversation (Carley et al. 2007). ORA also has general functionalities for data visualization and handling matrix operations associated with analyzing social networks.

3 Results

Our findings characterize the Twitter conversation surrounding NATO TRJE 2018 as primarily dominated by official NATO accounts. Notable information operations, however, also took place using bot and bot-like accounts. In the succeeding sections, we organize our results as follows: (a) an analysis of broad topics of discussion related to NATO Trident Juncture on Twitter (see Sect. 3.1); (b) an aggregated analysis of bot accounts triangulated with analysis of their locations and topics (see Sect. 3.2); (c) an individual-level drill-down that qualitatively characterizes the activities of ORA-identified influencers (see Sect. 3.3); and (d) a network-level drill-down of accounts which interacted with Sputnik accounts, which are known accounts linked to Russian media campaigns (see Sect. 3.4).

3.1 Topic analysis

Six topics were chosen for LDA based on coherence scores and manual assessment of the derived topics. Multidimensional scaling suggested that the last three topics

are relatively similar so they are interpreted concurrently in the succeeding analysis for convenience. Every tweet was assigned to a topic based on LDA topic probabilities. Figure 2 depicts the scaled distance between final topics used as well as their diffusion over time. Table 1 provides representative words and a sample tweet for each topic selected based on LDA results. We note that while the first two topics have coherent word lists, the latter two are relatively difficult to decipher on their own. This may be attributed to the fact that the topics themselves remain internally mixed.

Topic 1: NATO Trident Juncture Accounting for about 70% of all tweets (including retweets), the first topic appears to concern general updates about the NATO TRJE. Messages in this topic are characteristically descriptive of exercise activities. The primary topic therefore coincides with intuitive expectations of the tweets collected about the NATO Trident Juncture. Some simply document concrete happenings during the joint exercises. Others, as demonstrated by the sample tweet in Table 1, express solidarity with the efforts of NATO and alignment with its goals.

Topic 2: collision of the Helge Ingstad Accounting for about 2–3% of all tweets (including retweets), the second topic aggregates tweets that discussed naval vessels. Importantly, it captures the tweets mentioning the Helge Ingstad collision. As shown in the diffusion plot, most of the content in this topic was produced during the period directly following the crash. In the sample tweet presented, the crash of the Helge Ingstad is used to discredit the strength of NATO. Creatively mocking the Trident Juncture as a Trident Puncture, the account expresses disdain for participants in the exercise, implying they are incompetent and unworthy of respect.

Topic 3: world politics Comprised of tweets discussing general events around world politics, the third topic is not specific to the TRJE. It accounts for 6% of all tweets. Like the first topic, most of the conversation around this topic took place around the end of October and start of November, although there is also a spike at the start of the exercise. In the given example, NATO is mentioned as one of many entities linked to global conspiracies. While not immediately related to the military exercises, this topic appears to discuss NATO by embedding it in a broader context of secrecy and plots, hinting at a hidden global order of which NATO is merely a

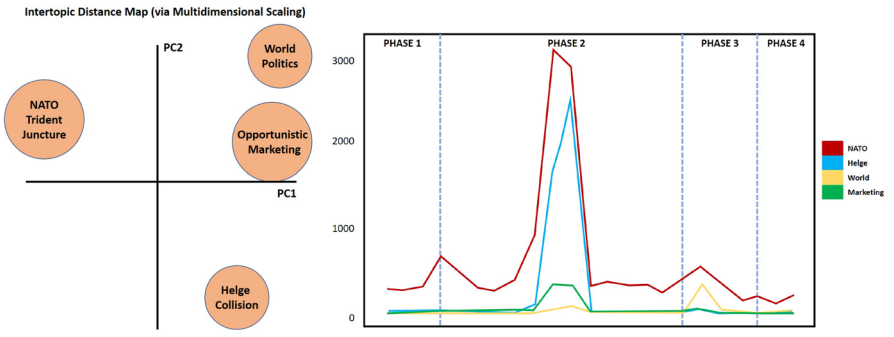

Fig. 2 Similarity and diffusion of topics. Left figure depicts multidimensional scaling of identified topics. Right figure depicts the relative prevalence of different topics over time

Table 1 Summary of topics identified using LDA. While some tweets specifically focused on NATO and events during the Trident Juncture Exercise, others focused more broadly on world politics or exploited public attention surrounding NATO to market various commodities online

Topics	Key words	Sample tweets
NATO Trident Juncture	Nato, exercise, trident juncture, norway, russia, military, maneuver, soldier, large, participate	JORSTADMOEN, Norway—A U.S. Army AH-64 Apache assigned to the 1st Battalion, 3rd Aviation Regiment. 12th Combat Aviation Brigade departs Rena Leir Airfield, Norway, during Exercise Trident Juncture, Nov. 5, 2018. TridentJuncture2018. WeAreNATO
Collision of Helge Ingstad	Russian, ship, photo, october, navy, frigate, sink, tanker, great, take	NATO has to learn from his TridentPUNCTURE. A fire on Canadian frigate HMCS Halifax. The storm nearly sank ship of USNavy GunstonHall. CanadianNavy ship #Toronto lost its turn. A collision with tanker, Navy frigate Norwegian KNM 'Helge Ingstad'
World politics	Go, prepare, need, medium, thank, want, leave, Ukraine, man, injure	There might be more truth here than you can handle... Just sayin' USA Syria Palestine MidEast Yemen Iraq Libya Afghanistan Kushner Zionism Nazi Wahhabism NATO NewWorldOrder UN AIPAC GenieEnergy Trump Obama Hillary GeorgeBush Obama BillClinton
Opportunistic marketing	Say, member, test, do, amp, defense, war, time, big	New on ebay: Arc Touch Wireless USB Receiver Mouse Slim Optical Flat Microsoft Touch Mouse KZ

part. Whether or not intentionally, such sentiments may be picked up in the online conversation due to the increased attention from the military exercises.

Topic 4: opportunistic marketing The final topic identified is an aggregation of three different topics with a high degree of overlap in terms. It accounts for 37% of original tweets and 20% of all tweets (when considering retweets). Again not primarily related to the NATO exercise, most of the conversation took place around the end of October and start of November (similar to the other topics), however it does not seem to increase activity during the start nor the accident. Other tweets found in this topic appear to insert the #nato hashtag without actually talking about NATO. Hashtags are used opportunistically just to boost the visibility of the commodities being marketed. This topic thus illustrates the numerous ways that the quick diffusion of information online is utilized for various purposes.

3.2 Bot analysis

Using a 60% threshold on Bothunter, we detected 24,686 bots in the dataset. This represents 30.27% of the unique users captured in our dataset. Cross-referenced against the role identity algorithm to remove special actors, only 10,072 bots remained, accounting for 12.35% of the unique users. Table 2 cross-tabulates the Bothunter results with role identity predictions. Percentages are relative to total users, with values in bold representing our final bot predictions.

The large reduction in final predictions is due to the significant proportion of reporters and government accounts classified by Bothunter as bot-like. Such Twitter users included the official NATO accounts which generated a high volume of content during the exercises. Figure 3 depicts the ORA visualization of bots on the social network of users connected by communication, with and without the role identity filtering. Filtered results are used as final predictions for conservative estimates of bot activity.

Cross-tabulated against our location prediction results, the top five countries with the highest numbers of detected bots included key NATO nations like the

Table 2 Summary of bot predictions cross-referenced against identity predictions

Role identity	Classified users		Detected bots	
	Accounts	%	Accounts	%
Regular users	28014	34.35	**10072**	**12.35**
Government	15914	19.51	9411	11.54
News agencies and reporters	10579	12.97	4456	5.46
Companies	905	1.11	308	0.38
Celebrities	906	1.11	233	0.29
Sports	551	0.68	206	0.25

For a conservative estimate of bot accounts, we use positively identified bots with Bothunter that our role identity algorithm does not classify as an otherwise special actor. Numbers in bold represent these finalized estimates

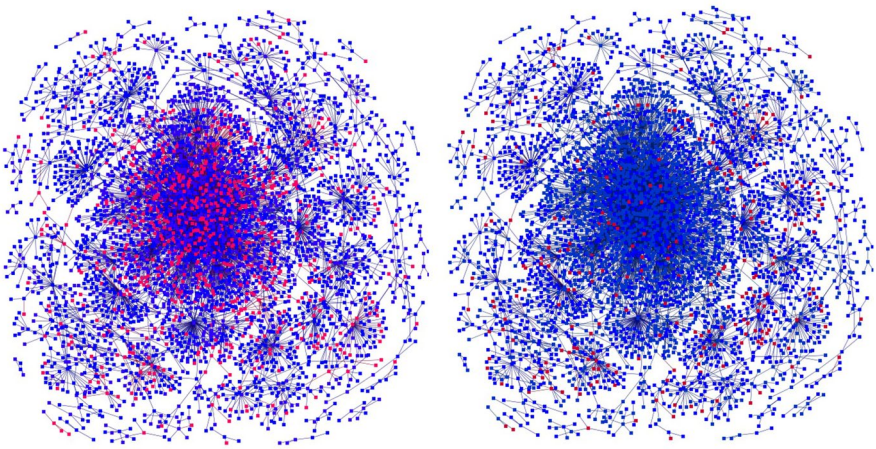

Fig. 3 Visualization of bots in the Twitter conversation surrounding the NATO Trident Juncture Exercise. Left figure depicts Bothunter-predicted bots in red and regular users as blue. Right figure depicts conservative bot estimates cross-referenced against role identity predictions

United States and Great Britain, with Russia also featuring a sizeable amount. Using filtered bot predictions, we report location results in Table 3. The share of bot accounts in each location suggests that while they are indeed prevalent in Russia, with which NATO presumably remains in notable tension, their activities are more focused on influencing the conversation in NATO states, as well as in Norway, where the exercises were held.

Finally, given the predicted bot accounts, we analyzed the levels of bot activity appearing in each of the conversation topics previously identified. As summarized by Table 4, bots appeared to feature most prominently in messages about the crash of the Helge Ingstad, followed by messages about NATO and the Trident Juncture exercises more generally. This suggests that, indeed, information operations were targeting the exercises, with the explicit goal of discrediting NATO's competence by focusing on the frigate collision.

Table 3 Predicted location of bot accounts

Location	Detected bots		Total
	Accounts	(%)	
United States	7956	38.15	20853
Great Britain	2716	36.65	7410
Norway	1865	35.39	5270
Russia	1293	19.40	6666
Spain	1203	16.32	7373

Percentages are in reference to the total number of accounts identified as belonging to the given location

Table 4 Summary of bot activity for each topic identified with LDA

Topic	Bot activity	
	Number of bot tweets	(%)
Collision of Helge Ingstad	2385	31.97
NATO Trident Juncture	42512	25.63
World politics	3018	20.30
Opportunistic marketing	3799	7.82

We organize rows by percentage of tweets in each topic associated with a predicted bot. The collision topic featured the highest level of predicted bot activity

3.3 Individual-level drill-down: influencer analysis

In this section, we present an individual-level drill-down of influential accounts as identified by ORA. Combining meta-network measures of centrality, ORA finds three types of influencers: Super Spreaders, Super Friends, and Other Influencers. Super Spreaders are users which generate content that is shared often, and hence spread information effectively. Super Friends are users that exhibit frequent two-way communication, facilitating large or strong communication networks. Other influencers are users which have an active network presence, by tweeting often or mentioning users often, and operate in central parts of the conversation, such as by using important hashtags, or mentioning important users. Our analysis showed that most influential accounts were NATO or otherwise government-affiliated, such as NATO, USNavy, and DeptofDefense. However, for unverified accounts, key influencers appeared to exhibit traits relevant to potential information operations. In the succeeding analysis, we characterize the activities of a sample of each type of influencer as identified by ORA. Usernames are withheld, however, for anonymity due to the sensitivity of content. These samples are summarized in Table 5

One exceptional user was classified as a Super Spreader after the crash of the Helge Ingstad. The account received a bot probability of 0.778. This score was likely obtained due to their extremely high Twitter activity (averaging 32 tweets/day for 5 years). Upon examination of their public profile, they appear to be a Russian patriot. In our dataset, the user calls the United States weak, citing that the NATO exercise could not be held without loss. Due to the coherence and awareness this tweet shows, it is unlikely to have been made by a bot. Further, the account exhibits intelligible direct replies to other users, in both English and Russian. From this

Table 5 Summary of influencers characterized for individual drill-down

Influencer type	Active phase	Bot Hunter score
Super spreader	After frigate collision	0.778
Super friend	During exercise	0.819
Other influencer	All phases	0.375

While NATO and government accounts were classified as influencers across all time phases, some unverified accounts were also identified as wielding significant influence

analysis it appears that they may be a cyborg account, a politically engaged citizen, or a paid user promoting Pro-Russian content.

Like the Super Spreader described above, the suspicious Super Friend is incredibly active on Twitter (averaging more than 55 tweets a day for 10 years, with increasing frequency). They claim to live in Russia, tweet anti-USA content, and responds coherently to other users. In our dataset, this user claims that Trump is Putin's puppet, and that he ordered Trump to surround Russia with troops. With so many of the user's tweets being replies or shared articles with detailed commentary, this user is also likely a paid tweeter or cyborg account.

This user is the most active user in the data set, with over 115 tweets per day on average for the last 10 years, with increasing frequency. Unlike the other suspicious users, this one claims to live in Sweden, backed up by profiles on other forms of social media linked in their Twitter account. Most of the tweets from the account are retweets, conveying pro-NATO/anti-Russian content. Even though most contents are retweets, the links to other social media profiles indicate that this account is likely not a bot. Hence, it seems most likely that this is an active, politically charged user.

3.4 Network-level drill-down: Sputnik analysis

In this section, we show how we may also incorporate prior knowledge about likely information operations into our interoperable pipeline. Following the suspicious activity geographically linked with Russia, we performed specialized analysis on the sub-networks associated with Sputnik accounts. Sputnik is a state-supported Russian media organization that has been previously linked to online propaganda. The succeeding analysis drills down on Sputnik activity in the following manner. In an initial step, all Sputnik Twitter accounts in the dataset are enumerated. All users with any form of communication (e.g., retweets, mentions) with these accounts are identified as first-order connections. Afterward, users communicating with first-order connections are included. The social network composed of this subset of users ($N = 6905$) is then subjected to analysis.

In terms of topics, about 82.21% of the tweets in the Sputnik sub-network discussed the main NATO hashtags. Drilling further down to the level of individual tweets, Sputnik stories included diverse types of anti-NATO messages and stories. Such stories revolved around themes of how TRJE aggravated local conflicts despite its ostensibly defensive objectives, triggering protests by peace activists around several Norwegian cities against the "NATO war". In one story, NATO troops had allegedly bruised and arrested an autistic man during drills. Stories had less than ten likes and retweets each.

Detected bot activity, however, was significant in the Sputnik sub-network. Despite their relatively low influence on the Twitter conversation, about 41.00% of users in the Sputnik sub-network were classified as bots, noticeably higher than the detected 12.35% throughout the entire dataset. ORA visualization is given in Fig. 4.

Overall, however, the relative influence of the Sputnik network was not very high, indicating that information operations were not particularly effective, at least those coursed through the known media channels. Measuring influence using Bonacich

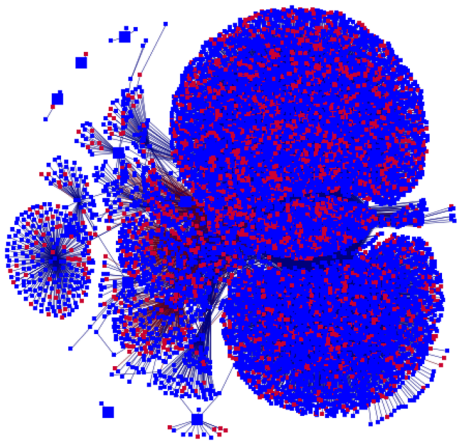

Fig. 4 The sub-network of accounts which interacted with Sputnik accounts. Red nodes represent accounts conservatively predicted to be bots. (Color figure online)

power centrality, a sociological network centrality measure of influence, Table 6 notes that NATO accounts wielded between × 5 and × 700 as much influence as Sputnik accounts over all time phases of the dataset.

4 Discussion

Overall, our integrated analysis offers a multi-scale view of bot activity surrounding the NATO Trident Juncture Exercises. We determined that bot activity was certainly non-negligible throughout the exercises, featuring significant activity in identified geographic locations, in relation to specific topics of conversation, and in association with known state-driven media. The bot activity we identified in this work appeared to seek to discredit the NATO alliance and its show of military force, painting NATO as brutal, incompetent, or unwanted by the general public. Such messages appeared to be targeted toward nations central to NATO, such as the US and the UK, as well as Norway, the host country for the exercises. The most influential agents driving these messages further appeared to be sophisticated cyborg accounts, not only generating a large volume of anti-NATO content through presumably automated messaging, but also expressing complex responses to other accounts suggesting the involvement of some human actors. We also captured the operations

Table 6 Bonacich power centrality of main Sputnik versus NATO account

Time phase	Sputnik	NATO	NATO/Sputnik
Before exercise	0.000001	0.000485	485
During exercise	0.000001	0.000682	682
Around crash	0.000006	0.000192	32
After crash	0.000036	0.000212	5.89

Throughout all four time phases, NATO accounts had significantly greater influence over the public conversation

of other bots which latched onto the publicity of the exercises, such as for opportunistic marketing and general conspiracy theories related to world politics.

These findings suggest that automated bots have indeed become a powerful and flexible tool utilized for a variety of purposes. Moreover, they demonstrate the value of computational analysis for sketching out the broad landscape of apparently multiple information operations and discerning how each may relate to various campaign objectives. Finally, we showed how prior awareness of Russian operations through Sputnik accounts can be readily integrated into various computational methods, thus allowing analysts to confirm or disconfirm pre-existing domain-specific hypotheses. Our proposed framework thus embodies social cyber-security's multidisciplinary nature, as it seeks to computationally enhance more politically, sociologically, and historically situated knowledge with large-scale empirical data.

To arrive at these insights, we employed an interoperable pipeline of computational tools to analyze different aspects of potential information operations surrounding the NATO Trident Juncture Exercise. Across multiple analytic junctures, our work demonstrates that methodologies integrating the principle of interoperability offer concrete benefits for analyzing online information operations. Firstly, utilizing multiple tools in a series allows for the accumulation of insight. Combining bot predictions with location predictions, for instance, allows analysis to map bot activity against geopolitical context, thereby elucidating relevant clues as to the objective of information operations under investigation. Combining insights of topic modeling and bot prediction also deepens these inferences, as analysts discover empirically which aspects of the online conversation feature significant participation (or perhaps lack thereof) of bots.

Secondly, utilizing multiple tools in parallel allows for the triangulation of insights. Algorithmic identification of bots can result in overprediction or underprediction depending on the relationship between the training dataset and the dataset encountered during deployment. By cross-referencing bot predictions with a second algorithm like role identification, we can generate more conservative estimates that avoid conflating automated bots and otherwise highly active Twitter users such as news agencies, celebrities, and government accounts. Drilled-down analysis of influencers identified via dynamic network analysis, when conducted in conjunction with Bothunter scores, likewise allows analysts to understand the degree to which bot accounts exert influence over the public conversation. In other cases, it may reveal new understandings of what kinds of accounts Bothunter may be classifying as bot (or bot-like), such as paid cyborg accounts.

Above all, interoperable pipelines broaden the possible questions with which analysts might engage datasets featuring potential information operations. As we demonstrated by focusing on the Sputnik subnetwork, a rich array of tools allows for greater flexibility in tackling different facets of, or communities involved in, information operations. By integrating principles of interoperability, analysts deploying methodological pipelines can arrive at a systematic understanding of the online conversation more comprehensive than the sum of individual, tool-level insights. As Tucker et al. (2018) show, rich social-scientific frameworks abound for theorizing disinformation campaigns and their impacts in online settings. Our methodological framework complements these conceptual developments by providing new ways of

analyzing data which go beyond traditional data sources in these disciplines (e.g., surveys, interviews) and arriving at large-scale, data-driven understandings of these phenomena that may enrich and be enriched by social theory.

Drawing on the broad framework we have laid out here, future work can integrate more tools that suit additional types of questions. For instance, sentiment analysis could be used to assess whether the messages being sent by bot and non-bot accounts were primarily positive or negative with respect to NATO and the exercises. Other types of dynamic network analysis, such as community detection or density measurement, could be used to study how different subnetworks evolve over time, or whether subnetworks associated with greater bot activity are likewise more echo-chamber-like. We also mention that our analysis was limited by several factors that future work may seek to address. The multilingual nature of the international conversation required a pre-processing step for translating all texts to English; although state-of-the-art translation models may indeed be highly performant, models of text in their original form may offer contextual nuances which translations do not effectively capture (Devlin et al. 2018). We also did not leverage other social media specific features like emojis or favorites, although such models have been developed with notable success.

Finally, we note that the analytic methods presented here are largely implemented on aggregated datasets. In a real-world setting, real-time detection and characterization of online information operations would allow stakeholders to respond in a more timely manner. Methods which can accommodate real-time streams of data would therefore be of significant value for succeeding developments in this field. Whereas data aggregation represents a non-negligible bottleneck to information operations analysis, however, the main steps in the interoperable pipeline may generally be run and replicated in the span of several hours, thus offering operationally relevant insights in relatively short amounts of time. This work thus showcases important problems, but also preliminary solutions, in conducting rapid detection and characterization of online information operations. Such improvements must develop alongside the evolution of legal, technical, and organizational limits of social media data usage for social cyber-security contexts.

While these methodological gaps do point to several areas for improvement in the pipeline we present here, they also illustrate the broad range of tools and analytic inquiries with which the principle of interoperable pipelines can be adapted for social cyber-security. Opportunities therefore abound for the development of interoperable methodologies specifically tailored for social cyber-security. Software attuned toward such pipelined integration of tools would likewise make a significant impact for deployment in this domain.

Acknowledgements This work is supported in part by the Office of Naval Research under the Multidisciplinary University Research Initiatives (MURI) Program award number N000141712675 Near Real Time Assessment of Emergent Complex Systems of Confederates, BotHunter award number N000141812108, award number N00014182106 Group Polarization in Social Media: An Effective Network Approach to Communicative Reach and Disinformation, and award number N000141712605 Developing Novel Socio-computational Methodologies to Analyze Multimedia-based Cyber Propaganda Campaigns. This work is also supported by the center for Computational Analysis of Social and Organizational 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. Additionally, Thomas Magelinski was supported by an ARCS foundation scholarship.

References

Al-Khateeb S, Hussain MN, Agarwal N (2019) Leveraging social network analysis and cyber forensics approaches to study cyber propaganda campaigns. Social networks and surveillance for society. Springer, Berlin, pp 19–42

Arif A, Stewart LG, Starbird K (2018) Acting the part: examining information operations within# blacklivesmatter discourse. In: Proceedings of the ACM on human–computer interaction 2(CSCW):20

Babcock M, Cox RAV, Kumar S (2019) Diffusion of pro-and anti-false information tweets: the black panther movie case. Comput Math Org Theory 25(1):72–84

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on web search and data mining, ACM, pp 65–74

Benigni M, Joseph K, Carley KM (2018) Mining online communities to inform strategic messaging: practical methods to identify community-level insights. Comput Math Org Theory 24(2):224–242

Bennett WL, Livingston S (2018) The disinformation order: disruptive communication and the decline of democratic institutions. Eur J Commun 33(2):122–139

Beskow DM, Carley KM (2018a) Bot conversations are different: leveraging network metrics for bot detection in twitter. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, pp 825–832

Beskow DM, Carley KM (2018b) Using random string classification to filter and annotate automated accounts. In: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 367–376

Beskow DM, Carley KM (2019a) Its all in a name: detecting and labeling bots by their name. Comput Math Org Theory 25:24–35

Beskow DM, Carley KM (2019b) Social cybersecurity: an emerging national security requirement. Mil Rev 99(2):117

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3(Jan):993–1022

Carley KM, Beskow DM (2017) Trident joust 2017, after action report. Technical report, Center for computational analysis of social and organizational systems, Carnegie Mellon University

Carley KM, Diesner J, Reminga J, Tsvetovat M (2007) Toward an interoperable dynamic network analysis toolkit. Decis Support Syst 43(4):1324–1347

Carley KM, Cervone G, Agarwal N, Liu H (2018) Social cyber-security. In: International conference on social computing. Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 389–394

Cheng J, Bernstein M, Danescu-Niculescu-Mizil C, Leskovec J (2017) Anyone can become a troll: causes of trolling behavior in online discussions. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, ACM, pp 1217–1230

Chew PA, Turnley JG (2017) Understanding Russian information operations using unsupervised multilingual topic modeling. In: International conference on social computing. Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 102–107

Conroy NJ, Rubin VL, Chen Y (2015) Automatic deception detection: methods for finding fake news. Proc Assoc Inf Sci Technol 52(1):1–4

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Dubois E, Gaffney D (2014) The multiple facets of influence: identifying political influentials and opinion leaders on twitter. Am Behav Sci 58(10):1260–1277

Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday. https://doi.org/10.2139/ssrn.2995809

Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59(7):96–104

Garrett RK (2009) Echo chambers online? Politically motivated selective exposure among internet news users. J Comput Mediat Commun 14(2):265–285
Huang B, Carley KM (2017) On predicting geolocation of tweets using convolutional neural networks. In: Lee D, Lin YR, Osgood N, Thomson R (eds) International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 281–291
Jin Z, Caro J, Zhang Y, Luo J (2016) News verification by exploiting conflicting social viewpoints in microblogs. In: Thirtieth AAAI conference on artificial intelligence
Karlsen R, Steen-Johnsen K, Wollebæk D, Enjolras B (2017) Echo chamber and trench warfare dynamics in online debates. Eur J Commun 32(3):257–273
Kudugunta S, Ferrara E (2018) Deep neural networks for bot detection. Inf Sci 467:312–322
Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D et al (2018) The science of fake news. Science 359(6380):1094–1096
Lee K, Caverlee J, Webb S (2010) Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval, ACM, pp 435–442
Lee K, Eoff BD, Caverlee J (2011) Seven months with the devils: a long-term study of content polluters on twitter. In: Fifth international AAAI conference on weblogs and social media
Mejias UA, Vokuev NE (2017) Disinformation and the media: the case of Russia and Ukraine. Media Cult Soc 39(7):1027–1042
Michelucci P, Shanley L, Dickinson J, Hirsh H (2015) A us research roadmap for human computation. arXiv preprint arXiv:150507096
Mihaylov T, Georgiev G, Nakov P (2015) Finding opinion manipulation trolls in news community forums. In: Proceedings of the nineteenth conference on computational natural language learning, pp 310–314
Montiel CJ, Boller AJ, Uyheng J, Espina EA (2019) Narrative congruence between populist president duterte and the filipino public: shifting global alliances from the United States to China. J Commun Appl Soc Psychol. https://doi.org/10.1002/casp.2411
Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In: Seventh international AAAI conference on weblogs and social media
Morstatter F, Wu L, Nazer TH, Carley KM, Liu H (2016) A new approach to bot detection: striking the balance between precision and recall. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), IEEE, pp 533–540
Nazer TH, Davis M, Karami M, Akoglu L, Koelle D, Liu H (2019) Bot detection: will focusing on recall cause overall performance deterioration? In: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 39–49
Nekmat E, Lee K (2018) Prosocial vs. trolling community on facebook: a comparative study of individual group communicative behaviors. Int J Commun 12:22
Qi S, AlKulaib L, Broniatowski DA (2018) Detecting and characterizing bot-like behavior on twitter. In: International conference on social computing. Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 228–232
Riquelme F, González-Cantergiani P (2016) Measuring user influence on twitter: a survey. Inf Process Manag 52(5):949–975
Röder M, Both A, Hinneburg A (2015) Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on web search and data mining, ACM, pp 399–408
Seah CW, Chieu HL, Chai KMA, Teow LN, Yeong LW (2015) Troll detection by domain-adapting sentiment analysis. In: 2015 18th international conference on information fusion (fusion), IEEE, pp 792–799
Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor Newslett 19(1):22–36
Stewart LG, Arif A, Starbird K (2018) Examining trolls and polarization with a retweet network. In: Proc. ACM WSDM, workshop on misinformation and misbehavior mining on the web
Tucker JA, Guess A, Barberá P, Vaccari C, Siegel A, Sanovich S, Stukal D, Nyhan B (2018) Social media, political polarization, and political disinformation: a review of the scientific literature. Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)

Uyheng J, Carley KM (2019) Characterizing bot networks on twitter: an empirical analysis of contentious issues in the asia-pacific. In: International conference on social computing. Behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, New York, pp 153–162

Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Eleventh international AAAI conference on web and social media

Wegner P (1996) Interoperability. ACM Comput Surv 28(1):285–287

Yang Z, Wang C, Zhang F, Zhang Y, Zhang H (2015) Emerging rumor identification for social media with hot topic detection. In: 2015 12th web information system and application conference (WISA), IEEE, pp 53–58

Zhou X, Zafarani R, Shu K, Liu H (2019) Fake news: fundamental theories, detection strategies and challenges. In: Proceedings of the twelfth ACM international conference on web search and data mining, ACM, pp 836–837

Joshua Uyheng is a Graduate Researcher with the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. His research interests include modeling polarization, disinformation, and collective emotions in the context of populist democracies.

Thomas Magelinski is a Graduate Researcher with the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. His research interests include dynamic network analysis and modeling legislative processes. He is supported by an ARCS Foundation Scholarship.

Ramon Villa-Cox is a Graduate Researcher with the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. His research interests include modeling interactional stance-taking and political polarization on social media.

Christine Sowa is a Graduate Researcher with the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. Her research interests include information dissemination and fake news on Reddit, focusing on adversarial takeovers of communities from foreign entities.

Kathleen M. Carley is Professor of Computer Science and Director of the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University and the CEO of Netanomics.